Dogs as Sentinels for Environmental Cancers

Addressing the Challenges of Spatial Epidemiology

Dissertation*

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Gianluca Boo

von

Bodio TI

Promotionskommission

Prof. Dr. **Sara I. Fabrikant** (Vorsitz)
Prof. Dr. **Andreas Pospischil**
Prof. Dr. **Kay W. Axhausen**
Prof. Dr. **Robert Weibel**
Prof. Dr. **Stefan Leyk**

Zürich, 2018

DOGS AS SENTINELS FOR ENVIRONMENTAL CANCERS

ADDRESSING THE CHALLENGES OF SPATIAL EPIDEMIOLOGY

Dissertation

Gianluca Boo

Department of Geography

University of Zurich

Committee

Prof. Dr. Sara I. Fabrikant, Prof. Dr. Andreas Pospischil,

Prof. Dr. Kay W. Axhausen, Prof. Dr. Robert Weibel, and Prof. Dr. Stefan Leyk

"Humility is indeed wise for the spatial analyst!"

Bailey and Gatrell 1995

ABSTRACT

Comparative research into canine and human cancers could help eradicate one of the primary causes of death globally. This is because the two species share both the same living environment and biological features involved in the development of cancer. In addition, given that the disease progresses much faster in dogs than humans, the geographic distribution of canine cancer can provide timely insights into the occurrence of shared environmental exposure, like, for instance, air pollution or radiation. Such a sentinel, or early-warning, application can benefit from the study methods of spatial epidemiology — a discipline examining the geographic variations in the distributions, determinants, and frequencies of diseases among populations.

However, when assessing statistical relationships between canine cancer and potential environmental exposures through the study methods of spatial epidemiology, a number of challenges may invalidate the resulting evidence. For this reason, this thesis aims to address four groups of challenges affecting the spatial epidemiology of canine cancer — (1) *society and context*, (2) *spatial data aggregation*, (3) *analytical framework*, and (4) *statistical inference*. These groups of challenges were examined through three case studies modeling canine cancer incidence rates retrieved from the Swiss Canine Cancer Registry (SCCR) data, the largest and longest-lived canine cancer registry to date assembled by the Collegium Helveticum, Zurich, for future comparative studies of canine and human cancers in Switzerland.

The first case study investigated the implications of underascertainment of cancer cases — a phenomenon that takes place when the diagnostic examination is not performed because of the dog's owner's decision. This data-quality challenge was dealt with by filtering out structural zeros from the incidence rates retrieved from the SCCR data because these sorts of zeros emanate from the sole absence of diagnostic examinations. Then, through model cross-validation, we evaluated the statistical relationships between canine cancer incidences and selected biologic risk factors as well as confounding variables accounting for underascertainment of cancer cases. This enabled determining the impact of specific societal and contextual settings on the quality of the SCCR data and the effects on the statistical performance and predictive power of the model.

Next, the second case study assessed the ubiquitous challenge of spatial data aggregation. This was via re-aggregating both the canine cancer incidence rates retrieved from the SCCR data and the explanatory variables using the extent of residential land within the municipal unit. Such a refinement step was inspired by the concept of dasymetric mapping — a cartographic technique designed to reflect the geographic distribution of the data more accurately. By contrasting models of canine cancer incidence rates based on dasymetrically refined and municipal units, we gained insights into changes in the coefficient estimates and statistical performance. These changes showed that dasymetric refinement could be employed to, at least partially, mitigate the effects of spatial data aggregation within the model.

At last, the third case study dealt with geographic variations in the statistical relationships estimated in models of average canine cancer incidence rates retrieved from the SCCR data by addressing the crucial issues of spatial non-stationarity and geographic scale. In doing so, we fit models across regions centered across every municipal unit of the study area and to varying spatial extents. Through value-by-alpha maps and scalograms, we exposed critical variations in the models' coefficients and performance as a function of the geographical location and scale of the regions. This highlighted the need to account for this condition by considering analytical frameworks enabling geographic variations of the coefficient estimations, for instance, through local or regional models.

The three case studies encapsulated the vital challenges and limitations of the spatial epidemiology of canine cancer that should be kept in mind for future studies with the SCCR data. Furthermore, the results facilitated understanding the general challenges involving statistical inference for potential environmental-sentinel applications, such as sample size, statistical power, and ecological fallacy. Following these findings, this thesis demonstrates that the spatial epidemiology of canine cancer can benefit from the conceptual and methodological framework of geography and, in particular, GIScience. With the potential of environmental-sentinel applications, further interdisciplinary research across the two disciplines is strongly advocated.

ZUSAMMENFASSUNG

Die vergleichende Erforschung von Krebserkrankungen bei Hunden und Menschen könnte dazu beitragen, eine der wichtigsten Todesursachen weltweit einzudämmen. Das liegt daran, dass die beiden Arten sowohl das gleiche Lebensumfeld als auch die gleichen biologischen Merkmale haben, die an der Entstehung von Krebs beteiligt sind. Da die Krankheit bei Hunden viel schneller voranschreitet als beim Menschen, kann die geografische Verteilung des Hundekrebses rechtzeitig Aufschluss über das Auftreten einer gemeinsamen Umweltbelastung, wie z.B. Luftverschmutzung oder Strahlung, geben. Eine solche Frühwarnung kann von den Methoden der räumlichen Epidemiologie profitieren - einer Disziplin, die die geographischen Unterschiede in den Verteilungen, Determinanten und Häufigkeiten von Krankheiten in der Bevölkerung untersucht.

Bei der Bewertung der statistischen Zusammenhänge zwischen Hundekrebs und möglichen Umweltexpositionen durch die Methoden der räumlichen Epidemiologie können jedoch eine Reihe von Herausforderungen die Beweiskraft von Ergebnissen schmälern. Aus diesem Grund untersucht diese Arbeit vier Gruppen von Herausforderungen, für die räumliche Epidemiologie von Hundekrebs: (1) *Gesellschaft und Kontext*, (2) *räumliche Datenaggregation*, (3) *analytischen Rahmen*, und (4) *statistische Inferenz*. Diese Gruppen von Herausforderungen wurden anhand von drei Fallstudien zur Modellierung der Inzidenzraten von Hundekrebs untersucht. Die zugrunde liegenden Daten der Studie stammen aus dem Schweizerischen Krebsregister (SCCR), mit Sitz in Zürich, dem bisher grössten und am längsten bestehenden Krebsregister des Collegium Helveticum für vergleichende Studien über Hunde- und Humankrebs in der Schweiz.

Die erste Fallstudie untersuchte die Auswirkungen der Unterbewertung von Krebserkrankungen - ein Phänomen, das dann auftritt, wenn die diagnostische Untersuchung aufgrund der Entscheidung des Hundebesitzers nicht durchgeführt wird. Eine der großen Herausforderung dieser Fallstudie war die Verbesserung der Datenqualität. Hierzu wurden strukturelle Nullen aus den Inzidenzraten der SCCR-Daten herausgefiltert, da diese Art von Nullen allein auf das Fehlen diagnostischer Untersuchungen zurückzuführen ist. Im Anschluss daran werteten wir die statistischen Beziehungen zwischen Hundekrebs Inzidenzen und ausgewählten biologischen Risikofaktoren mit

Modell Cross-Validierungen aus. In diesem Zusammenhang werteten wir auch verwirrende Variablen aus, die häufig für die Unterbewertung von Krebsfällen verantwortlich sind. Dadurch konnten die Auswirkungen bestimmter gesellschaftlicher und kontextbezogener Einstellungen auf die Qualität der SCCR-Daten und die Auswirkungen auf die statistische Leistung und Vorhersagekraft des Modells ermittelt werden.

In der zweiten Fallstudie wurde die allgegenwärtige Herausforderung der Geodatenaggregation untersucht. Dabei wurden sowohl die aus den SCCR-Daten gewonnenen Inzidenzraten für Hundekrebs als auch die erklärenden Variablen anhand der Ausdehnung von Wohnflächen innerhalb der Gemeindeeinheit neu aggregiert. Ein solcher Verfeinerungsschritt wurde durch das Konzept der dasymetrischen Kartierung inspiriert - eine kartographische Technik, die die geographische Verteilung der Daten genauer widerspiegelt. Durch die Gegenüberstellung von Modellen der Inzidenzraten von Hundekrebs, die auf dasymetrisch verfeinerten und kommunalen Einheiten basieren, konnten wir Erkenntnisse über Veränderungen in den Koeffizientenschätzungen und der statistischen Leistung gewinnen. Diese Änderungen zeigten, dass die dasymetrische Verfeinerung eingesetzt werden kann, um die Auswirkungen der räumlichen Datenaggregation innerhalb des Modells zumindest teilweise abzumildern.

Die dritte Fallstudie befasste sich schliesslich mit den geografischen Variationen in den statistischen Beziehungen, die in Modellen der durchschnittlichen Inzidenzraten von Hundekrebs aus den SCCR-Daten geschätzt wurden, indem sie die entscheidenden Fragen der räumlichen Nichtstationarität und der geografischen Skala behandelte. Dabei passen wir Modelle regionenübergreifend über alle kommunalen Einheiten des Untersuchungsgebietes und in unterschiedlichen räumlichen Ausmassen an. Mit Hilfe von Value-by-Alpha-Karten und Skalogrammen haben wir kritische Schwankungen der Koeffizienten und der Leistung der Modelle in Abhängigkeit von der geographischen Lage und dem Massstab der Regionen aufgezeigt. Daraus ergab sich die Notwendigkeit analytische Rahmenbedingungen zu berücksichtigen, die geografische Variationen der Koeffizientenschätzungen ermöglichen, beispielsweise durch lokale oder regionale Modelle.

Die drei Fallstudien fassen die wesentlichen Herausforderungen und Grenzen der räumlichen Epidemiologie von Hundekrebs zusammen, die für

zukünftige Studien mit den SCCR-Daten berücksichtigt werden sollten. Darüber hinaus ermöglichen die Ergebnisse ein erleichtertes Verständnis der allgemeinen Herausforderungen, die mit statistischen Schlussfolgerungen für potenzielle Anwendungen von Umwelt-Wächtern verbunden sind, wie z.B. Stichprobengrösse, statistische Aussagekraft und ökologische Täuschung. Die Ergebnisse dieser Arbeit zeigen, dass die räumliche Epidemiologie von Hundekrebs kann von den konzeptionellen und methodischen Rahmen der Geographie und insbesondere GIScience. Mit dem Potenzial von Environmental-Sentinel-Anwendungen wird eine weitere interdisziplinäre Forschung über beide Disziplinen hinweg stark gefördert.

ACKNOWLEDGEMENTS

This thesis was carried out at the Department of Geography of the University of Zurich (UZH) and the Collegium Helveticum Zurich, Swiss Federal Institute of Technology and University of Zurich (ETHZ/UZH) and would not have been possible without their infrastructure and funding.

Within these institutions, I would like to thank my main Ph.D. supervisors Prof. Dr. Sara I. Fabrikant (Department of Geography, UZH), Prof. Dr. Andreas Pospischil (Collegium Helveticum, ETHZ/UZH), and Prof. Dr. Kay W. Axhausen (Collegium Helveticum, ETHZ/UZH; Transport Planning, ETHZ) for assisting me throughout my endeavors. Special thanks goes to Prof. Dr. Stefan Leyk (Department of Geography, University of Colorado Boulder) for his patient supervision during the design and implementation of the case studies developed as part of this thesis. I also want to thank Prof. Dr. Robert Weibel (Department of Geography, UZH) for the much-appreciated contribution to the final wrap-up. Apart from my Ph.D. committee members, I thank Prof. Dr. Christopher Brunsdon (National Centre for Geocomputation, Maynooth University) for his inputs to the third case study and Prof. Dr. Eric Delmelle (Department of Geography and Hearth Sciences, University of Charlotte) for acting as an external reviewer.

I further extend my gratitude to my colleagues at the Department of Geography (UZH), but I do not mention any names because I am afraid I may forget someone. Nevertheless, special acknowledgments belong to Sascha Credé, Ekaterina Egorova, Ismini Lokka, Sara Maggi, and Jérémie Ory (borrowed from the French Institut National de l'Information Géographique et Forestière – IGN) for their support. Also, without attempting to name everyone, I thank my colleagues at the Collegium Helveticum Zurich (ETHZ/UZH), especially Dr. Ramona Graf and Dr. Katrin Grüntzig, who have preprocessed and assembled the Swiss Canine Cancer Registry (SCCR). I also acknowledge the far-too-often underestimated work of the editorial boards and anonymous reviewers involved with the journals where I submitted my case studies.

Lastly, I thank the entire Boo family – Mariantonietta, Flavio, Gabriele, Mary, Giulia, Ivàn, and Hagos – and my good friends from Lausanne – Duccio, Filippo, Giulia, Gregory, Lauro and Raffaella, Max, Patrick, Samuel, and Sebastiano. Their support and friendship helped me to keep focused on

the preparation of this thesis and, when necessary, to forget about it. Above all, I thank my wife Osnat for her unconditioned love and valuable help with the graphic design of this thesis and its figures.

This thesis is dedicated to my son, Adam Aviv, and his sister, Mia, and to the memory of my grandmothers, Vinicia and Sandra.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC – Akaike information criterion

CAnCER – Guelph Companion Animal Cancer Epidemiologic Registry

CANR – California Animal Neoplasm Registry

CAR – Conditional autoregressive model

CHF – Swiss Francs

E(S)DA – Exploratory (spatial) data analysis

ETS – Environmental tobacco smoke

FCI – Fédération Cynologique Internationale

FOPH – Swiss Federal Office of Public Health

GAM – Geographical Analysis Machine

GIS – Geographic information systems

GIScience – Geographic Information Science

GLM – Generalized linear model

GWR – Geographically weighted regression

ICD-O – International Classification of Diseases for Oncology

IQR – Interquartile range

LISA – Local indicators of spatial association

MAE – Mean absolute error

MWR – Moving window regression

MAUP – Modifiable areal unit problem

NCI – United States National Cancer Institute

NICER – Swiss National Institute for Cancer Epidemiology and Registration

PCOP – Purdue Comparative Oncology Program

RCA-SP – Sao Paulo Animal Cancer Registry

RBD – Swiss Federal Register of Buildings and Dwellings

RL – Relative likelihood

RMSE – Root-mean-square error

SCCR – Swiss Canine Cancer Registry

SAR – Simultaneous autoregressive model

SFOT – Swiss Federal Office of Topography (swisstopo)

SFSO – Swiss Federal Statistical Office

SFTA – Swiss Federal Tax Administration

SMR – Standard morbidity (or mortality) ratio

US – United States of America

WHO – World Health Organization

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Fighting the war on cancer

On December 23, 1971, the President of the United States (US) – Richard Nixon – signed the National Cancer Act. This momentous event, commonly referred as day zero of the "war on cancer," marked the starting point of an unprecedented effort to eradicate one of the primary causes of death globally (Sporn 1996). The massive investments and initiatives promoted by the US National Cancer Institute (NCI) over the following decades drove national and international commitments to research and drug development.

As brilliantly summarized by De Vita and Rosenberg (2012), these endeavors resulted in significant advances both in the understanding of the nature of the disease and in the treatment of various forms of it. However, despite these achievements, the World Health Organization (WHO) (2017) attributed 8.2 million deaths to cancer in 2012, and 14.1 million new cases were estimated for the same year. Moreover, these numbers are expected to rise by approximatively 70% over the next two decades as populations grow, age, and adopt lifestyle behaviors that increase cancer risk (Vineis and Wild 2014).

Such an alarming trend suggests that the war on cancer cannot stand only on a three-legged stool consisting of radiation therapy, surgery, and chemotherapy, but also needs to consider public health policies of primary prevention. The reason being is that fighting cancer through primary prevention could address risk factors, such as unhealthy behaviors and environmental exposure (Torre et al. 2016). According to recent estimates of the WHO (2017), tackling these risk factors could reduce up to half the current global burden of the disease, thus, becoming one of the major battles in this war.

To develop policies for cancer primary prevention, national and international public health agencies rely on, among other elements, the

evidence provided by epidemiologic investigations (De Vita and Rosenberg 2012). Epidemiology is a discipline concerned with the "study of the distribution and determinants of disease frequency in man" (MacMahon and Pugh 1970). Distribution, determinants, and frequency, the three components of this definition, indicate that diseases are not random phenomena.

For this reason, epidemiologic investigations decompose disease occurrence through the three-sided analytical prism of person, time, and place (Ahrens et al. 2005). When the focus is on the place where the disease occurs, a specific study approach ought to be considered — spatial epidemiology. This approach is located at the intersection between the disciplines of epidemiology and geography and seeks to understand the geographic variations in the distributions, determinants, and frequencies of diseases among different populations (Elliott et al. 1996).

Spatial epidemiology encompasses three broad types of study methods — disease mapping, disease clustering, and geographic correlation studies. Disease mapping depicts the geographic distribution of the disease frequencies across populations within a study area. Then, disease clustering determines whether the disease frequencies are spatially dispersed, randomly distributed, or spatially aggregated. Lastly, geographic correlation studies test relationships between disease frequencies, distributions, and potential determinants, or, in other words, the risk factors (Lawson 2006).

These three study methods have, in several instances, provided evidence concerning risk factors well in advance of other analytical approaches, such as laboratory studies. As such, the spatial epidemiology of cancer is rapidly gathering interest from researchers via the study of several risk factors, especially environmental exposures, over the last decades (Roquette et al. 2017).

1.1.2. Investigating environmental exposures

To develop and test hypotheses regarding statistical associations between cancer and environmental exposures, spatial epidemiology is contingent on the availability of information retrieved from diagnostic cases (Boscoe et al. 2004). In the spatial epidemiology of cancer, this information is usually stored in population-based registries. These databases contain diagnostic

data on the individual cancer cases for a population of known size and composition (Parkin 2008). However, owing to privacy considerations, individually identifiable information is typically not available for research (Boscoe et al. 2004).

This sensitive information consists of demographic characteristics, employment history, healthcare plans, and residential addresses of the individuals that have been deceased for less than 50 years (Gliklich et al. 2014). Given these privacy considerations, spatial epidemiology addresses the statistical associations between cancer and environmental exposures at the population level, or stated another way, for aggregated groups of individuals rather than for the individuals themselves. Cancer cases are enumerated, for instance, as a function of the residential address at the time of diagnosis within geographic units, such as administrative districts or ZIP codes (Boscoe et al. 2004).

The process of spatial data aggregation features several kinds of challenges when investigating the relationships between enumerated cancer cases and environmental exposures. These are based on, for example, the inability of isolating concurrent exposures occurring within residential, occupational, and recreational settings. Moreover, given the relatively long latency period of many cancer types and the increased frequency of migratory movements, it is often difficult to determine which residential location presents the environmental exposure of interest (Ward and Wartenberg 2006).

To control for these sources of exposure misclassification, the spatial epidemiology of cancer could make use of companion animals as models of environmental cancers in humans (Schmidt 2009; Reif 2011). Similar to humans, diagnostic information on individual disease cases is stored in companion animal cancer registries. These registries consist of databases compiled by the histopathologic units of veterinary hospitals (Brønden et al. 2007). Dogs, in particular, benefit from a relatively high level of veterinary care, resulting in detailed information on individual cancer cases.

Most importantly, information on the demographic characteristics and residential addresses of dogs is more readily available for researchers because

of lower privacy restrictions. Besides data quality and accessibility, compared to other companion animals, including birds, felines, and rodents, dogs are excellent models of cancer development in humans. The reason for that is the striking biological similarities between the two species, which leads to an increased predisposition to cancer in certain dog breeds (Rowell et al. 2011). Another critical feature of dogs is that they intimately share the household with their owners. Thus, changes in the frequency of canine cancer can inform the study of environmental exposures within residential settings.

In this regard, the spatial epidemiology of canine cancer can also enable a more accurate assessment of shared environmental exposures given that the disease progresses more rapidly in dogs than in humans. Shorter latency periods can also allow transferring the evidence into a context of early detection of exposures associated with human cancer, or, in other words, using dogs as sentinels for cancers associated with environmental exposures (Schmidt 2009; Reif 2011).

1.1.3. Using dogs as environmental sentinels

Employing companion animals as sentinels for environmental exposures resulting in human health outcomes is not new (Reif 2011). An iconic illustration is the canary that, in the 19th century, accompanied miners during the extraction of coal. Based on a higher susceptibility to carbon monoxide poisoning, the canary was employed to detect the quality of air in mining tunnels. If a canary would faint and fall from its perch, miners would quickly exit the tunnel in search of oxygen (Schmidt 2009).

This example shows that the primary goal of considering dogs as sentinels for environmental cancers is to provide early warnings on a critical exposure before this can affect humans. As shown in **Figure 1**, for such a purpose, it is vital to choose an appropriate study approach to link cancer in dogs to environmental exposures, potentially affecting humans (Rabinowitz et al. 2009; Scotch et al. 2009). To produce this sort of evidence, it is necessary to investigate relationships at the population level, for instance, between enumerated canine cancer cases and shared environmental exposures.

When examining such relationships in geographic correlation studies, two classes of challenges may invalidate the provided evidence. The first class

concerns the epidemiological nature of the investigation, where the most-often reported issues are different forms of bias and confounding factors arising, for instance, from diagnostic errors. These potential limitations are well known and routinely considered in the spatial epidemiology both of human and canine cancers (Bartlett et al. 2010; Elliott and Wakefield 2000). The second class of challenges concerns the geographic nature of the investigation, where issues associated with spatial data quality, aggregation, and analysis are commonly understood with the spatial epidemiology of human cancer (Elliott and Wartenberg 2004; Jacquez 2004). Nevertheless, these challenges have not been systematically addressed in the spatial epidemiology of canine cancer to date.

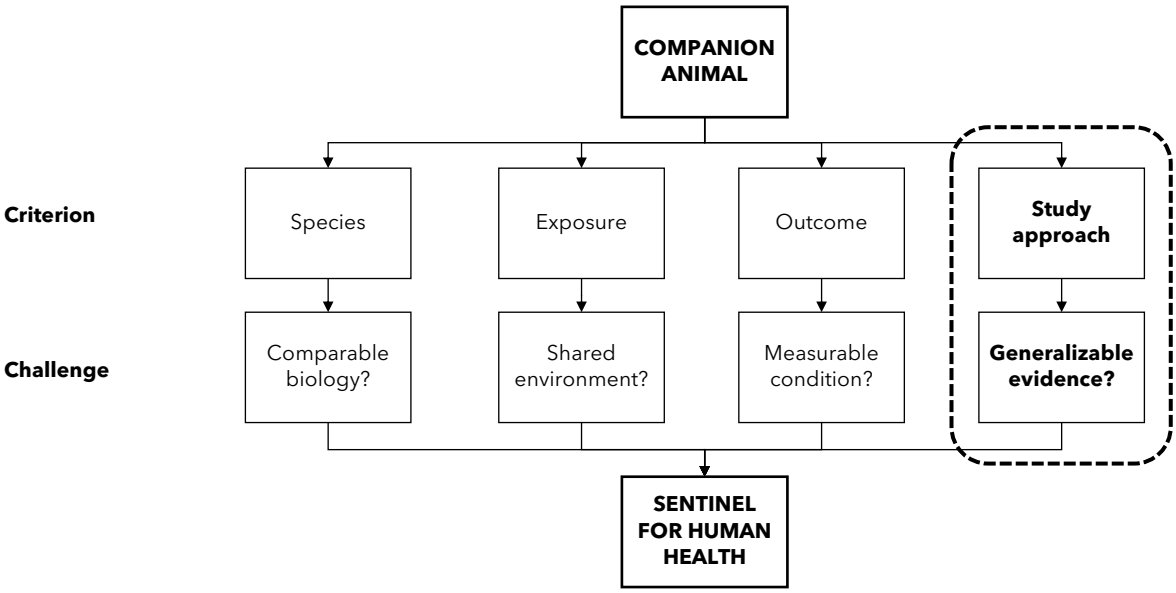

Figure 1. Companion animals as sentinels for human health – criteria and challenges. The challenges and limitations of the spatial epidemiological approach (dashed line) need to be carefully addressed to provide generalizable evidence for potential environmental-sentinel applications. The figure is adapted from the Canary Database (2016).

In this regard, Geographic Information Science (GIScience) can provide insights into the geographic nature of spatial epidemiology (Raubal et al. 2013). This connection between the two disciplines is witnessed in the seminal methodological reviews, among others, by Elliott et al. (1996, 2000), Glass (2000), Rezaeian et al. (2007), Boulos (2004), Meade and Emch (2010), and Páez et al. (2015), as well as in extensive discussions on the challenges of spatial epidemiology (Jacquez 2000, 2004; Elliott and Wartenberg 2004;

Beale et al. 2008). In particular, Jacquez (2004) proposed a taxonomy consisting of four groups of challenges and limitations of the spatial epidemiology of human cancer. In detail:

(1) Challenges and limitations imposed by specific *societal and contextual* settings, which influence the individual perceptions both of the disease and its relationships with potential environmental exposures (Elliott and Wartenberg 2004).

(2) Challenges and limitations imposed by the *spatial data*, which mostly connect with the process of spatial data aggregation of individual cancer cases and the resulting location and attribute uncertainty (Boscoe et al. 2004).

(3) Challenges and limitations imposed by the selected *analytical framework*, as insufficient knowledge of the data can result in models that poorly estimate the statistical associations between cancer and potential environmental exposures (Roquette et al. 2017).

(4) Challenges and limitations regarding *statistical inference*, which cannot be directly derived from observed geographic patterns or statistical associations, among others, because of generalization or statistical power considerations (Jacquez 2004).

Motivated by the potential environmental-sentinel applications enabled by the spatial epidemiology of canine cancer, this thesis aims to address these challenges and limitations.

1.2. OBJECTIVE AND RESEARCH QUESTIONS

1.2.1. Research questions

Following the motivations presented earlier, the objective of this thesis is to tackle the challenges and limitations of the spatial epidemiology of canine cancer for potential environmental-sentinel applications. Owing to the taxonomy provided by Jaquez (2004) and the knowledge gaps identified in Section 2.4, this thesis seeks to answer the research questions (RQ) presented hereafter.

RQ 1 How does *society and context* challenge the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

RQ 2 What are the implications of using *spatial data* in the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

RQ 3 How does the selected *analytical framework* impact the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

RQ 4 How does the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors impact *statistical inference* for potential environmental-sentinel applications?

As these research questions directly concern the geographic nature of spatial epidemiology, this thesis also connects to specific knowledge areas of GIScience.

1.2.2. Relevance to GIScience

This thesis has the objective of dealing with challenges and limitations of the spatial epidemiology of canine cancer for potential environmental-sentinel applications. However, given that the focus of the investigation is on the geographic nature of spatial epidemiology, the specific contributions also relate to several thematic areas of the body of knowledge of GIScience. This is because GIScience is a discipline that goes beyond the mere storage,

analytics, and visualization capabilities of Geographic Information Systems (GIS), as it uses them as tools for scientific understanding (Clark 1997).

Such a basic but expressive definition, which has been more recently updated, among others, by Mark and Turk (2003) and Goodchild (2008), suggests that this thesis connects to several knowledge areas of GIScience. As portrayed in **Figure 2**, the body of knowledge of GIScience, originally summarized by DeBiase et al. (2006) in the seminal publication, "Geographic Information Science & Technology Body of Knowledge," currently consists of 10 broad thematic areas (DiBiase et al. 2017). Three of those thematic knowledge areas are relevant for interpreting the specific contributions of this thesis within the discipline of GIScience.

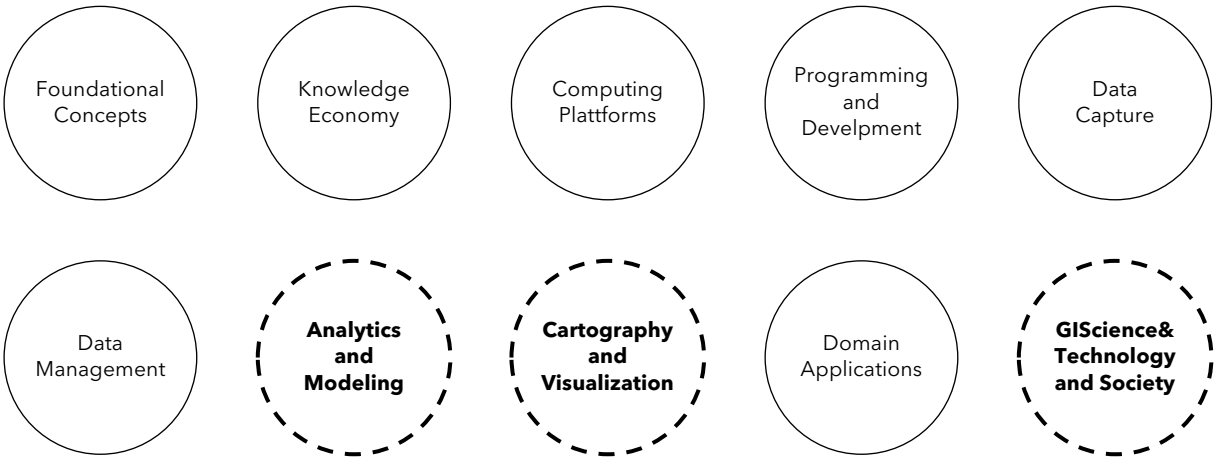

Figure 2. The 10 thematic knowledge areas included in the latest version of the "Geographic Information Science & Technology Body of Knowledge" by DiBiase et al. (2017). This thesis aims to contribute to the body of knowledge via the areas underscored by a bolded dashed line. The figure is adapted from DiBiase et al. (2017).

As indicated in **Figure 2**, the three thematic knowledge areas (KA) connected with this thesis are briefly introduced hereafter with explicit reference to the research questions (RQ) presented earlier, in Section 1.2.1.

KA 1 The theme "Analytics and Modeling" is broadly concerned with the creation of knowledge on the geographic processes and their distributions (DiBiase et al. 2017). This knowledge area is addressed throughout the thesis by investigating the challenges and limitations associated with *society and context* (RQ 1), *spatial data aggregation* (RQ 2), and the *analytical framework* (RQ 3).

KA 2 The theme "Cartography and Visualization" focuses on the general design and use of maps and mapping technology (DiBiase et al. 2017). This knowledge area is confronted by assessing the challenges and limitations associated, in particular, with the issues of *spatial data aggregation* (**RQ 2**) and spatial structure connected with the *analytical framework* (**RQ 3**).

KA 3 The theme "GIScience & Technology and Society" relates to the different impacts of GIScience, from the institution down to the individual level (DiBiase et al. 2017). This knowledge area is addressed by evaluating the challenges and limitations associated with *society and context* (**RQ 1**) as well as with *statistical inference* (**RQ 4**) for potential environmental-sentinel applications.

To better understand the connections between GIScience and the spatial epidemiology of canine cancer, the contributions of this thesis will also be discussed with respect to these three thematic knowledge areas.

1.3. WORKFLOW AND STRUCTURE

1.3.1. Research workflow

The four groups of challenges and limitations highlighted by Jacquez (2004) are still widely unacknowledged in the spatial epidemiology of canine cancer at present. Therefore, the research workflow adopted in this thesis was developed in a classical exploratory fashion (Shields and Rangarajan 2013). In essence, this approach involves selecting specific challenges and limitations known to affect the spatial epidemiology of human cancer, and testing to which extent they also affect the spatial epidemiology of canine cancer.

For such an exploratory purpose, the primary data source examined in this thesis was the Swiss Canine Cancer Registry (SCCR) – the largest and longest-lived canine cancer registry to date (Grüntzig et al. 2015, 2016). In particular, the research questions on *society and context* (**RQ 1**), *spatial data* (**RQ 2**), and *analytical framework* (**RQ 3**) were addressed in three distinct case studies of canine cancer frequencies retrieved from the SCCR during the period 2008-2013. At the time of writing, these case studies have been submitted and partially published in relevant peer-reviewed scientific journals. In detail:

RQ 1 is addressed in Boo et al. (2017) through a case study of the *implications of underascertainment of cancer cases*, a phenomenon occurring when the dog owner would not seek diagnosis and treatment for canine cancer. The introduction of structural zeros in the computation of canine cancer incidence was expected to impact the statistical performance of models of canine cancer incidence;

RQ 2 is addressed in Boo et al. (2018b) through a case study of the *effects of spatial data aggregation*, namely enumerating canine cancer cases within municipal units. The process of spatial data aggregation was expected to affect the estimation of statistical associations in models of canine cancer incidence rates when including independent variables explicitly related to density and distance; and lastly

RQ 3 is addressed in Boo et al. (2018a) through a case study of the *influences of spatial non-stationarity and geographic scale* in models of average canine cancer incidence rates. These were anticipated to be critical to the statistical performance of the model and the estimation of the statistical associations, because of important spatial heterogeneity in the geographic characteristics of the study area.

To avoid self-plagiarism and extensive self-citation, references to the three case studies are provided at the beginning of each relevant section. The case studies are also examined to answer the last research question, namely regarding *statistical inference* (**RQ 4**) for potential environmental-sentinel applications. As part of the "One Medicine – One Oncology" research project (Pospischil et al. 2015), the work presented in this thesis is also related to co-authored work in veterinary epidemiology. That which is published in peer-reviewed scientific journals is systematically cited to acknowledge the important contributions of the other authors.

An exhaustive list of featured publications developed as part of the "One Medicine – One Oncology" research project is presented in the annexes in Section 7.4.

1.3.2. Structure of the thesis

This thesis consists of seven chapters, including this introductory chapter, **Chapter 1**. The remainder of this thesis is structured as follows.

Chapter 2 outlines the conceptual background of the thesis by reviewing the relevant scientific literature on the different approaches, challenges, and limitations of spatial epidemiology. The literature review focuses on applications in the spatial epidemiology of human cancer to highlight their inherent challenges and limitations. This chapter also discusses relevant canine cancer data sources and the few existing comparative studies of canine and human cancers.

Chapter 3 presents the three case studies in distinct sections named according to the thematic contribution – (**1**) *implications of underascertainment of cancer cases*, after Boo et al. (2017); (**2**) *effects of*

spatial data aggregation, after Boo et al. (2018b); and (3) *influences of spatial non-stationarity and geographic scale*, after Boo et al. (2018a). These sections describe in detail the materials, methods, results, and key findings for the related case studies.

Chapter 4 presents a general discussion of the results of the three case studies, with reference to the identified knowledge gaps. It also presents the general limitations of the research project, as linked with epidemiological and statistical modeling considerations. The previous discussions are considered for answering the proposed research questions. Finally, the connections between this thesis and the body of knowledge of GIScience are also investigated.

Chapter 5 concludes by summarizing the specific achievements concerning the proposed research goals and provides an outlook on directions for future work. **Chapter 6** lists the references that supported this thesis, and **Chapter 7** includes different annexes, for example, reference maps and a list of pertinent knowledge areas of GIScience.

2. LITERATURE REVIEW

This chapter reviews the relevant scientific literature to define the conceptual and methodological framework of this thesis. The first section details the three study methods of spatial epidemiology – disease mapping, cluster analysis, and geographic correlation studies – concentrating on research of human cancer. The second section discusses their potential limitations within the taxonomy proposed by Jacquez (2004). The third section reviews existing canine cancer data sources and comparative studies.

2.1. APPROACHES OF SPATIAL EPIDEMIOLOGY

2.1.1. Disease mapping

Disease mapping is the oldest and perhaps most common study method of spatial epidemiology (Walter 2000). In essence, this study approach is constituted by the cartographic representation of the geographic distribution of disease frequencies, for instance, incidence or prevalence, within selected populations. A crucial element in disease mapping is modeling the disease diagnostic cases as spatial objects (Lawson et al. 2001). For this purpose, the most direct way is to model the disease according to residential location, typically as a point feature.

Although point-based disease maps accurately indicate the presence of risk factors, considering the privacy considerations presented earlier, disease cases often need to be spatially aggregated within spatial units (Beale et al., 2008; Elliott and Wartenberg, 2004). This process typically involves the enumeration of the new cases recorded within each spatial unit during a specific period, in other words mapping the disease incidence. To account for the distribution of the new disease cases across the different at-risk populations, disease incidence can also be computed as rates conditioned to the underlying at-risk populations recorded during the same period, thus mapping the disease incidence rates (Lawson et al. 2001).

To further refine the distribution of the new cases within the different at-risk populations, incidence rates can also be adjusted – or standardized – based on demographic characteristics, such as age, sex and ethnicity, as well as other confounding factors. This adjustment not only enables a more accurate cartographic representation of the observed new cases, but also an estimation of the expected number of new cases. By calculating a ratio between the observed and the estimated new cases, it is finally possible to map disease risk as a standard morbidity (or mortality) ratio (SMR) (Lawson et al. 2001).

Disease incidence, disease incidence rates, and disease risk can also be smoothed according to some assumed spatial and statistic distribution in the data, to account for spurious spatial patterns due to random variations in the observed new cases. This result is achieved through various analytical frameworks, which are not directly considered in this thesis but extensively discussed, among others, by Best et al. (2005), Lawson (2006), and Lawson et al. (2001).

While disease maps were already in existence during the 17th century, the most cited example is the map of the London cholera epidemic of 1854 produced by Snow (1855). Snow's seminal work was closely followed by the publication of several cartographic representations of the geographic distribution of cholera and other diseases that have been extensively reviewed elsewhere (Walter 2000; Lawson et al. 2001; Koch 2017). During the same period, Haviland (1855) created the first set of maps of cancer mortality for the north of England. However, just 70 years later, cancer prevalence began to be consistently mapped in England and Wales through the tremendous effort of Stock (1928, 1936, 1937, 1939).

More extensive work involving cancer mapping was carried out in the 1940s following the progressive institution of national programs of cancer registration (Boyle et al. 2012; Beam 2013). For example, the first atlases of cancer incidence were produced in Scandinavia for the period between 1943 and 1980 (Møller Jensen et al. 1988), and in England and Wales for the period between 1968 and 1985 (Swerdlow and Silva 1992). The development of early computers substantially benefitted the efforts of cancer mapping, especially

starting from the 1970s onwards, when Burbank (1971) published the first computer-drawn map of cancer mortality in the US.

In the following decades, the increasing availability of data from cancer registries, together with methodological and technological advances in the domain of GIS, resulted in an unprecedented milestone in cancer mapping. These activities were extensively reviewed, among others, by Boyle et al. (2012) and d'Onforio et al. (2016), wherein it was highlighted that, over the past 50 years, cancer mapping has permitted the identification of previously unknown risk factors for several cancer types. For example, the United States Cancer Atlas by Mason et al. (1975) allowed for the subsequent association of oral cancer to the use of chewing tobacco in the seminal study by Blot and Fraumeni (1977).

These factors all suggest that cancer mapping is a first, important, step in generating hypotheses on associated risk factors that need to be tested in follow-up studies, such as disease clustering and geographic correlation studies (Greenberg 1985).

2.1.2. Disease clustering

Disease clustering extends the insights provided by disease mapping, by testing whether the observed geographic patterns in the disease incidence and rates are because of random fluctuations or reflect the true variations in the frequency of a disease (Kulldorff and Nagarwalla 1995). Although there is a broad consensus on the definition of disease cluster provided by Knox (1989) – "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance," the operationalization of disease clustering is often widely debated (Wakefield et al. 2000).

Given the purpose of this thesis, this section only discusses the most common study methods involving the detection of clusters of disease incidence, rates, and risk, especially concerning spatial dependence, in other words, spatial autocorrelation. More extensive reviews of disease clustering, including different approaches, such as clustering and cluster investigation, can be found elsewhere (Alexander and Boyle 1997; Wakefield et al. 2000; Waller and Gotway 2004; Lawson 2006). The most popular study method for

detecting clusters of disease incidence, rates, and risk are distance/adjacency methods and moving-window-based methods (Wakefield et al. 2000).

The first method facilitates detecting spatial autocorrelation globally, for example, through Moran's I (Moran 1948), Geary's c (Geary 1954), and Getis-Ord General G (Cliff and Ord 1973) statistics. Spatial autocorrelation can also be assessed locally, for example, using the Getis-Ord Gi* (Getis and Ord 1992) statistic, and the local indicators of spatial association (LISA) (Anselin 1995). In essence, testing for global spatial autocorrelation establishes whether disease incidence, rates, and risk are spatially dispersed, randomly distributed, or clustered (Moran 1950). Testing for local spatial autocorrelation enables detecting where similar values are spatially autocorrelated. Local clusters of higher disease incidence, rates and risk are identified as disease "hotspots" (Anselin 1995).

A second method for detecting disease hotspots consists of determining the significance of disease incidence, rates, and risk falling into a window of fixed size (Wakefield et al. 2000). This approach is primarily related to two distinct analytical methods. The Geographical Analysis Machine (GAM), the first scan method, which is founded upon a distance-based window (Openshaw et al. 1987). The second and most common scan method is SatSCAN, which utilizes a population-size-based window, typically of circular shape (Kulldorff et al. 1997).

Investigating clusters of disease incidence, rates, and risk has become an increasingly popular approach in the spatial epidemiology of cancer over the last decades (Roquette et al. 2017). Compared to cancer mapping, this study method was developed relatively recently. In fact, one of the first sets of studies on cancer clustering was probably carried out by Cruickshank (1940, 1947), which – following Stock's work (1928, 1936, 1937, 1939) – investigated clusters of cancer mortality in England and Wales employing a crude measure of autocorrelation.

Almost 30 years later, Glick (1979) marked the beginning of studies on the global spatial autocorrelation of cancer with a case study of cancer mortality in the US state of Pennsylvania. This groundbreaking investigation examined the effects of various adjacency levels (i.e., spatial lags) on the

results of the Moran's I test (Glick 1979). Since then, sensitivity tests evaluating different distance/neighboring levels for different cluster-detection methods have been systematically adopted in the detection of cancer clusters (Wakefield et al. 2000). Among countless examples, Walter et al. (1994) developed an influential study on the clustering of the incidence of different cancer types in the province of Ontario, Canada, scrutinizing the results of the Moran's I test and a non-parametric rank adjacency statistic for a variety of weights.

More recently, local spatial autocorrelation tests have also been adopted in the detection of cancer hotspots, for instance, across countries of Western Europe (Rosenberg et al. 1999), in regional districts in Taiwan (Tsai and Perng 2011), in local communities in Saudi Arabia (Al-Ahmadi and Al-Zahrani 2013), and in neighborhoods within the city of Shenzhen, China (Zhou et al. 2015). Earlier studies of cancer hotspots mostly featured moving-window methods. For example, Openshaw et al. (1988) detected local clusters of leukemia in specific municipalities in Northern England using the GAM, and Kulldorff et al. (1997) observed local clusters of breast cancer in the northeast of the US using SatSCAN.

In spite of the availability of alternative disease-clustering approaches, such as cluster-investigation methods, the cluster-detection studies presented before are considered as an essential stepping stone for corroborating preliminary assumptions on the geographic patterns of cancer developed through disease mapping (Wakefield et al. 2000). Finally, these hypotheses are tested in geographic correlation studies.

2.1.3. Geographic correlation studies

The third study method of spatial epidemiology is geographic correlation studies (Elliott et al. 1996). This approach aims to investigate the geographic distribution of diseases, by evaluating statistical associations to risk factors, such as environmental exposures, socioeconomic and demographic characteristics, and lifestyle factors (Waller and Gotway 2004).

To be tested in a statistical framework, the independent variables associated with these risk factors must be computed within the same analytical units employed for the aggregation of disease cases (Gulliver et al.

2015). To date, different frameworks allow for testing statistical associations between diseases and related risk factors. These are extensively discussed by, among others, Elliott et al. (2000) and Lawson (2006). This section exclusively covers analytical frameworks to test for statistical associations involving incidence and rates of relatively rare diseases, such as cancer.

For this purpose, the most frequently utilized method is perhaps the generalized linear model (GLM), when assuming the incidence data follows a Poisson distribution (Frome 1983; Frome and Checkoway 1985). However, as the incidence data may be overdispersed (Berk and MacDonald 2008), alternative models – such as the negative binomial, zero-inflated, and hurdle models (Hu et al. 2011; He et al. 2014) – can also be fit in geographic correlation studies. These models can be generated both within a frequentist and Bayesian framework.

Furthermore, the geographic distribution of the incidence data may be spatially autocorrelated (Wall 2004) and/or the statistical associations to the relevant risk factors may not be constant across space, thus exhibiting spatial non-stationarity (Fotheringham et al. 1996). In the first case, the statistical model has to account for spatial autocorrelation, typically through empirical Bayes applications, like the simultaneous autoregressive model (SAR) (Whittle 1954) or the conditional autoregressive model (CAR) (Besag 1974). More recently, a fully Bayesian analysis of the CAR model has also been developed (Lunn et al. 2009).

In the second case, several analytical frameworks embed spatial non-stationarity, typically through local models (Fotheringham and Brunsdon 1999; Leung et al. 2000). In geographic correlation studies, the most common local model is the geographically weighted regression (GWR) (Lloyd 2010). It was originally developed by Fotheringham et al. (1996) to understand spatial non-stationarity within statistical associations.

Among the earliest geographic correlation studies of cancer, Apperly (1941) was the first to suggest relationships between skin cancer mortality and solar radiation across various states of the US by directly associating the two variables through scatterplots. The same approach was later applied in several works developed, for instance, by Breslow and Enstrom (1974). In this

study, associations between cancer mortality and the consumption of alcohol and tobacco in the US were demonstrated (Breslow and Enstrom 1974). During the same period, White (1972) was one of the pioneers of employing a Poisson regression to assess the geographic distribution of leukemia mortality in England and Wales.

The development of the Cancer Mortality Atlas by Manson et al. (1976) led to a series of formal and informal geographic correlation studies testing hypotheses on associated risk factors, including environmental exposures, specific activities, and behaviors (Elliott and Wartenberg 2004). Among these studies, the most important ones are by Hoover et al. (1975), Mason (1976), Blot and Fraumeni (1977), and Blot and Fraumeni (1982). As mentioned earlier, Blot and Fraumeni (1977) were the first to link oral cancer to the use of chewing tobacco.

Beginning in the 1980s, geographic correlation studies were also carried out on smaller scales (Richardson 1996). These studies were mostly performed in the US, by linking different cancer types to the characteristics of drinking water (Carpenter and Beresford 1986), exposure to radiation (Edling et al. 1982), and pesticides (Stokes and Brace 1988). In the 1990s, geographic correlation studies started to account for spatial autocorrelation through CAR and SAR models, for instance, in the works of Mollié (1990) and Mollié and Richardson (1991).

More recently, commencing in the 2000s, geographic correlation studies based on local models, such as GWR, have been employed to assess heterogeneous statistical associations between cancer and associated risk factors across study areas. Among these studies, cervical cancer risk was found to be related to socioeconomic status in England (Cheng et al. 2011), and cancer risk was linked with air toxins in the US state of Florida (Gilbert and Chakraborty 2011). As noted by Roquette et al. (2017), geographic correlation studies permit testing the hypotheses on associated risk factors developed through disease mapping and clustering, and their popularity in the spatial epidemiology of cancer is steadily rising.

The spatial epidemiology study methods feature several challenges and limitations. These are described in detail in the next section according to the taxonomy proposed by Jacquez (2004).

2.2. CHALLENGES OF SPATIAL EPIDEMIOLOGY

2.2.1. Society and context

Society and context are considered to play a significant role in the success of disease mapping, cluster analysis, and geographic correlation studies, especially for primary prevention purposes (Best and Wakefield 1999; Jacquez 2004; Elliott and Wartenberg 2004). In this regard, several authors have highlighted effects on false positives (Wartenberg 1999; Elliott and Wakefield 2000; Jacquez 2004), like, for example "false alarms" resulting from reports of potential disease clusters around pollution sources compiled by groups of concerned citizens (Kulldorff et al. 1998) or from informal geographic correlation studies conducted on the basis of personal concern (Elliott and Wartenberg 2004).

For a primary prevention purpose, false negatives have by far a greater impact because they may impede detecting risk factors associated with the disease (Wartenberg 1999; Elliott and Wakefield 2000). These false negatives can arise because of, for instance, underestimation resulting from the incompleteness of the incidence data retrieved from cancer registries (McNamee 2003; Chen et al. 2014; Thygesen and Ersbøll 2014). Incompleteness is a manifestation of two distinct and often confused phenomena – underreporting and underascertainment of disease cases (Gibbons et al. 2014). The former occurs at the healthcare level when the result of a performed diagnostic examination is not reported in a disease registry because of incorrect diagnosis or inaccurate notification (Gibbons et al. 2014). The latter takes place at an earlier stage when the individual is not seeking a diagnostic examination, and, as a consequence, it is not performed at all (Gibbons et al. 2014).

The underreporting of disease cases can be considered a contextual limitation in the spatial epidemiology of cancer as it may vary geographically as a function of the healthcare system, service, and practitioner (Teppo et al. 1994). Moreover, the underascertainment of cancer cases can result from both contextual and societal influences on the individual use of healthcare services – this is the typical cause of a non-diagnosis (Sethi et al. 1999). Fiscella et al. (2000) reported that disparities in the general use of healthcare

services could be linked to several factors, for instance, affordability (Potosky et al. 1998), geographic access (Perloff et al. 1997), education level (Pincus et al. 1998), knowledge (Brown et al. 1990), and health beliefs (Lannin et al. 1998).

A crucial aspect connected with the underreporting of cancer cases is that the diagnostic examination is performed. Therefore, a simple comparison between the cases recorded within the healthcare system and the registry can allow estimating the degree of underreporting (Teppo et al. 1994; Thygesen and Ersbøll 2014). Another related method consists of comparing the observed number of cases in the registry with an expected number of cancer cases from a standard population derived from a similar data source (Goldberg et al. 1980; Thygesen and Ersbøll 2014). Lastly, perhaps the most commonly employed and critiqued approach is the capture-recapture method, which estimates the sensitivity of two independent samples based on cancer cases reported in both instances (Robles et al. 1988; Schouten et al. 1994).

Assessing underascertainment of cancer cases is by far more challenging because the diagnostic examination has not been performed at all. For this reason, the completeness of the cancer registry data cannot be directly estimated through the same imputation methods presented earlier. Besides, as made note of by Heckman (1979), regardless of the estimation method, this issue cannot be entirely overcome. However, given the factors resulting in disparities in the use of healthcare services reported by Fiscella et al. (2000), the underascertainment of disease cases can be considered as a systematic source of error resulting from varying types of selection bias (Hernán et al. 2004). Such a form of bias is introduced by the non-random selection of individuals, in this specific case, being the undergoing of a diagnostic examination. Extreme manifestations of underascertainment of disease cases may lead to extremely unreliable disease-registry data (Elliott and Wakefield 2000; Beale et al. 2008).

A relatively simple insight into the degree of selection bias can be gained by examining the significance and strength of correlations between the retrieved cancer incidence and rates as well as confounding factors

associated with selection bias, such as socioeconomic, demographic, and healthcare characteristics (Clegg et al. 2009). Furthermore, as introduced by Heckman's influential work on sample selection bias (Heckman 1974, 1976, 1979), these individual explanatory factors can also be included as independent variables. For instance, these can be embedded in a simple GLM (Wirth and Tchetgen 2014) or more complex statistical frameworks, such as decision trees (Gibbons et al. 2014) and Bayesian models (Dvorzak and Wagner 2016). Besides these specific issues affecting data quality, spatial epidemiology also faces more general challenges associated with the use of spatial data.

2.2.2. Spatial data

The challenges described previously compromise the quality of the incidence data retrieved from cancer registries (Jacquez 2004). Still, another critical matter relates to the spatial character of the disease data. In this regard, a primary concern is location (or positional) uncertainty – it is assumed that the geographic distribution of disease cases can be approximated through residential coordinates or addresses (Beale et al. 2008). This exercise relies on the questionable assumption that locational attributes are accurate and that the place of residence is a meaningful surrogate of the exposure (Jacquez and Jacquez 1997; Jacquez and Waller 2000). To cope with this limitation, the process of spatial data aggregation can somehow mitigate these and other sources of uncertainty (Roquette et al. 2017).

Still, it is essential that the selected unit of aggregation is somehow representative of the individual mobility patterns while considering the exposure and the latency period (Jacquez 2004). Taking into account these considerations, aggregating individual cancer cases within spatial units is a common practice in the spatial epidemiology of cancer (Roquette et al. 2017). Yet, spatial data aggregation has several potential issues, often grouped together under the term of the modifiable areal unit problem (MAUP) (Openshaw 1984). This fundamental concept was first described by Gehlke and Biehl (1934) and later formalized by Openshaw and Taylor (1979) and Openshaw (1984). Specifically, it involves the influence of both the shape

and scale of the spatial unit in the computation of summary values, such as counts, rates, proportions, and densities.

Several studies have demonstrated the impact of the MAUP in different applications of spatial epidemiology, such as disease mapping (Walter and Birnie 1991; Morris and Munasinghe 1993; Choi et al. 2003), cluster detection (Waller and Turnbull 1993; Olson et al. 2006; Ozonoff et al. 2007), and geographic correlation studies (Cleek 1979; Fotheringham and Wong 1991; Holt et al. 2010). Especially in geographic correlation studies, the computation of multiple independent variables involving several types of spatial support – such as point, line, areal units, and surfaces – can be particularly challenging because of various MAUP effects along with the possible incompatibility across the available spatial supports (Cressie 1996; Cressie et al. 2009).

As a general rule, spatial epidemiological studies involving areal data are carried out at the finest scale possible, with subsequent testing for changes occurring at coarser scales (Roquette et al. 2017). However, particularly in geographic correlation studies, the processes of data measurement and collection can produce spatial variables computed using different reference units. These may also have incompatible levels of granularity, for instance, across census tracts, ZIP codes, and administrative boundaries (Gotway and Young 2002).

To cope with the problems associated with such changes of spatial units, several geostatistical solutions exist. These have been extensively reviewed by, among others, Lawson (2006), Gotway and Young (2002), and Cressie (1996). When dealing with areal data, most of these solutions are centered around the concept of areal interpolation, or in other words, the process of transferring spatial data from a spatial unit into another (Gotway and Young 2002). Well-known areal interpolation methods are the inverse-distance interpolation and kriging in their different variants (Waller and Gotway 2004).

Much earlier in terms of general timelines, another solution to cope with the effects of the MAUP was provided by a cartographic technique, namely dasymetric mapping (Eicher and Brewer 2001; Mennis and Hultgren 2006). This cartographic technique was created in the 1920s by Semenov-Tian-Shansky (Petrov 2012), further developed by Wright (1936), and, more

recently, by a host of others. The aim is to produce more accurate geographic distributions of areal data considering geographic context utilizing some related ancillary spatial variables (Eicher and Brewer 2001).

Both the areal interpolation and dasymetric mapping methods permit producing compatible spatial variables at the finest scale possible. However, studying the multiple effects of spatial data aggregation at different scales is key to understanding complex processes, such as those examined in spatial epidemiology (Graham et al. 2004; Banerjee et al. 2014). This can be achieved through statistical methods for multi-scale processes, such as, for instance, what was suggested in the work of Tobler (1989), that, together with Cressie (1996) and Fotheringham (1989), recommended choosing models with parameters that change predictively across scales to cope with the MAUP. Among these models, multiscale spatial tree models, such as the ones proposed by Basseville et al. (1992) and Chou et al. (1994), are perhaps the most representative, as they allow predicting processes occurring at different scales. The complex issues associated with spatial data highlight the importance of choosing an appropriate analytical framework.

2.2.3. Analytical framework

Given the limitations presented earlier, a crucial aspect of disease mapping, cluster detection, and geographic correlation studies is choosing an appropriate analytical framework (Elliott and Wartenberg 2004; Beale et al. 2008). In many instances, this implies that both the variables associated with the disease incidence (or rates) and risk factors, as well as the parameters required by the selected analytical framework, should be correctly specified (Jacquez 2004).

With the different approaches of spatial epidemiology, the most critical parameters are perhaps the ones formalizing the spatial structure of the areal data, specifically the relationship of spatial dependency between observations, being disease incidence, rates, and associated risk factors (Bavaud 1998; Brunsdon et al. 2002). Regardless of the specific terminology, these parameters are computed based on a spatial-weight matrix – a bi-dimensional expression of the spatial dependence between observations (Anselin 1995). In a spatial-weight matrix, the diagonal has only zeros while

the off-diagonal elements take either one (i.e., connected) or zero (i.e., non-connected) values, and are often row-standardized accordingly – scaled to sum up to one (Getis and Ord 1992; Bivand 1998).

The two most common approaches to defining the off-diagonal elements of a spatial-weight matrix for areal data are the distance- and neighborhood-based methods (Anselin 1988; Bivand 1998). The former builds a matrix based on the Euclidean distance between the centers or centroids of each spatial unit, while the latter examines the neighborhood structure, for instance, accounting for the adjacency of units sharing a common border (Anselin 1988; Bivand 1998). While the properties of these two methods are still widely debated, there is a consensus in the somewhat vague statement that "good candidates must reflect the properties of the particular phenomena" (Bavaud 1998).

Another crucial element to defining a weight matrix is choosing a threshold value that determines which spatial data is connected (i.e., one) or non-connected (i.e., zero) to another one (Cliff and Ord 1973). This procedure is comprised of defining a general threshold distance between the centroid of the different spatial units or a general threshold number of nearest neighbors. Furthermore, universal threshold values may not be valid for the entire study area and depend upon their relative location, hence requiring more flexible solutions (Foster and Gorr 1986).

The primary issue for choosing an analytical framework is the knowledge of the underlying spatial system (Jacquez 2004; Lawson 2006). With this in mind, Jacquez (2004) and many others have wisely promoted the use of exploratory (spatial) data analysis (E(S)DA) methods to provide a basic understanding of spatial and statistical associations. Among these methods, disease mapping and cluster detection are considered the most instructive (Lawson 2006). While disease mapping involves the visual examination of the spatial structure in the data, cluster detection consists of performing sensitivity tests based on different distance or neighboring threshold values for various cluster-detection methods (Wakefield et al. 2000).

However, when the exploratory approach seeks to understand the role of spatial structure in the relationships between disease incidence, rates, and

one or more independent variables, local models, such as the moving window regression (MWR) and GWR, need to be considered (Fotheringham and Brunsdon 1999; Lloyd 2010). These models enable evaluating spatial structure through geographic variations in the model parameters. In general, while smaller bandwidths yield larger variation in the model parameters, greater bandwidths result in minor variations (Fotheringham et al. 2003).

When employing GWR, there are various techniques for selecting an appropriate bandwidth, for instance, through cross-validation, or, more commonly, the Akaike information criterion (AIC) (Akaike 1974; Browne 2000). These techniques permit determining the best bandwidth for estimating the spatial-weight function — usually in the form of a fixed or adaptive kernel — that maximizes the statistical performance of the local model. A well-known example of a spatial-weight function is, for instance, the Gaussian kernel (Lloyd 2010). However, within a study area, there may be regions where a more localized bandwidth leads to better statistical performance than others.

This issue involves the concept of geographic scale, or in other words, the local spatial extent employed for parameter estimation (Lloyd 2014). If the local model parameters differ considerably across a range of geographic scales, it is highly problematic to determine the scale at which the process operates (Tate and Atkinson 2001). Conversely, the persistence of similar model parameters across different geographic scales facilitates identifying independent variables for the process of interest (Jacquez 2004). Although these considerations can involve selecting an appropriate analytical framework, statistical inference may have inherent limitations.

2.2.4. Statistical inference

Spatial epidemiology enables testing whether the geographic distribution of diseases and relationships to risk factors are statistically significant (Jacquez 2004). In spite of the advantages provided by the different hypothesis-testing frameworks, the resulting statistical inference may have inherent limitations, as heavily elaborated upon, among others, by Elliott and Wartenberg (2004), Jacquez (2004), and Waller and Gotway (2004).

Given the purpose of this thesis, three distinct, but often confused, limitations of statistical inference are reviewed, namely, generalizability, ecological fallacy, and statistical power (Guthrie and Sheppard 2001). Generalizability is a vital challenge for spatial epidemiology and involves statistical inference drawn from a study sample that is transferred to a target population or other populations (St. Sauver et al. 2012). Generalizable — or externally valid — findings are mostly contingent upon their relevance and the internal validity of the study in question, therefore involving appropriate study design, data collection, and analytical frameworks (Kukull and Ganguli 2012).

A logical limitation in the quest for generalizability is the ecological fallacy (Lawson 2006). This issue involves statistical inference drawn at the group level, which is wrongly transferred to the individuals belonging to the groups (Piantadosi et al. 1988). The MAUP, for instance, can be considered a geographic manifestation of the ecological fallacy as statistical inference based on spatial units should not transfer at the individual level (Openshaw 1984). To prevent such a mistake, spatial epidemiology studies involving areal units should be considered as purely hypothesis-generating, because statistical inference at the individual level cannot be directly transferred (Guthrie and Sheppard 2001).

The last challenge, erroneously thought to be a matter of generalizability, is statistical power (Kukull and Ganguli 2012). This arises when proceeding to formal hypothesis testing, in terms of statistical significance, for statistical inference. In essence, this concept relates to the likelihood that the test will detect as statistically significant an effect that is indeed present (Kukull and Ganguli 2012). Higher statistical power features a higher probability of successfully identifying an effect while a lower statistical power produces the opposite result, namely a type II error (Lawson 2006).

In geographic correlation studies, statistical power is mostly contingent upon effect and sample size, because larger effects and samples typically result in a higher statistical power (Cohen 1992). With respect to the objective of this thesis, additional considerations involving statistical power in disease clustering are not reviewed, but are extensively discussed elsewhere (Elliott and Wartenberg 2004; Waller et al. 2006).

The key to estimating the internal validity of research in spatial epidemiology lies in the assessment of potential changes in the performance of the selected analytical framework (Kukull and Ganguli 2012). As mentioned earlier, in cluster detection, this is carried out, for instance, through sensitivity tests. A more systematic method that may apply to geographic correlation studies is cross-validation.

Cross-validation incorporates a set of techniques to assess the internal and external validity of a statistical model (Snee 1977; Cattin 1980). Ultimately, the different cross-validation techniques are based on the use of a subset of the sampled data to train a statistical model (i.e., the training set), then testing it on another subset (i.e., the validation set). A necessary condition for conducting cross-validation is that both the training and the validation sets are sampled from the same population (Browne 2000; Steyerberg et al. 2001).

There are two particular methods of cross-validation — exhaustive and non-exhaustive cross-validation. While the former tests all the possible ways to divide the original sampled data into training and validation sets, the latter does not (Browne 2000). To establish the internal validity and the potential external validity of a statistical model, cross-validation facilitates the computation of various measures of the predictive power, like, for example, the mean average error (MAE) and the root mean squared error (RMSE) (Picard and Cook 1984). Such measures directly inform about the internal validity of a model, but, besides its relevance, the definition of external validity is dependent upon two potential issues.

The first issue relates to non-stationary temporal trends in the sampled data, which makes it impossible to consider the training and validation sets as part of the same population (Hyndman and Koehler 2006). The second issue is based on model misspecification, which can dramatically inflate measures of the predictive power of certain validation sets (Akaike 1974). The size of the training and validation sets is an additional consideration for cross-validation because of statistical power. In his notable work, Cohen (1988, 1992, 1995) proposed a relatively straightforward way to determine a minimum sample

size resulting in a suitable statistical power, which can typically apply in most geographic correlation studies.

Spatial epidemiology can also be applied to studying canine cancer. The available data sources and findings from comparative studies of human and canine cancer are reported in the next section.

2.3. SPATIAL EPIDEMIOLOGY OF CANINE CANCER

2.3.1. Available data sources

While the first human cancer registries date from the 1940s, the earliest animal cancer registries were established only twenty years later, in the 1960s (Brønden et al. 2007). In the early stages of animal cancer registration, three canine cancer registries are particularly important – the California Animal Neoplasm Registry (CANR), US (active between 1963 and 1966) (Dorn 1967), the Kansas University Neoplasm Registry, US (active between 1961 and 1971) (Strafuss 1976), and the Tulsa Registry of Canine and Feline Neoplasms, US (active between 1972 and 1977) (MacVean et al. 1978).

Since the end of the 1970s, several population-based animal cancer registries involving dogs have been initiated, and most of them are still currently active – the Purdue Comparative Oncology Program (PCOP), US (active since 1979) (Lengerich et al. 1992), the Animal Tumor Registry of Genoa province, Italy (active between 1985 and 2002) (Merlo et al. 2008), the Norwegian Cancer Project (active since 1990) (Gamlem et al. 2008), the VetCancer Registry, worldwide (active since 1994) (Brønden et al. 2007), the Registry of Canine Tumours in Sweden (Agria Pet Insurance) (active since 1995) (Bonnett and Egenvall 2010), the Piedmont Canine Cancer Registry, Italy (active between 2001 and 2008) (Baioni et al. 2017), the Danish Veterinary Cancer Registry (active since 2005) (Brønden et al. 2010), the Animal Tumour Registry of the Venice and Vicenza provinces, Italy (active since 2009) (Vascellari et al. 2009), the Guelph Companion Animal Cancer Epidemiologic Registry (CAnCER), Canada (active since 2010) (Nødtvedt et al. 2011), the Sao Paulo Animal Cancer Registry (RCA-SP), Brazil (active since 2013) (Tedardi et al. 2015), and the Swiss Canine Cancer Registry (SCCR) (currently covering the period between 1955 and 2014) (Grüntzig et al. 2015).

The SCCR was built based on diagnostic data provided by the Vetsuisse institutes of veterinary pathology in Berne and Zurich and a private diagnostic laboratory located in the Zurich area (Grüntzig et al. 2015, 2016). The diagnostic data were harmonized and combined in a FileMaker database and exported as a comma-separated tabular file, following a procedure described by Grüntzig and colleagues (2015). In this tabular file, the residential address

was reported in the form of the postcode, which was subsequently linked to the unique identification number of the Swiss municipal units. This was done using a correspondence table available on the Swiss Federal Statistical Office website (SFSO 2017) as reference index to join these identifying attributes. Such a basic geocoding step enabled for the spatial aggregation of the individual diagnostic attributes within municipal units. By this means, for instance, canine cancer incidence could be linked to ancillary spatial and non-spatial data available at the municipal level, such as municipal boundaries (SFOT 2017) or official statistics (SFSO 2017).

Besides the canine cancer registries presented above, many other extemporary data sources have been assembled for spatial and purely epidemiology studies of canine cancer. For example, a dataset was retrospectively collected from diagnostic records issued from seven out of 21 French veterinary histopathologic laboratories between 2001 and 2002 (Pastor et al. 2009). In addition, less extensive data collection efforts are routinely carried out within the histopathologic unit of most veterinary hospitals. For example, Reif (1995) studied 93 cases of canine lymphoma at the Colorado State University Veterinary Teaching Hospital (collected between 1987 and 1990), and later another 103 cases of canine nasal cancer (collected between 1986 and 1990) (Reif et al. 1998).

Compared to the almost 500 human cancer registries surveyed by Parkin (2006) a decade ago, existing canine cancer registries are fewer in number and often short-lived. Besides the issue of data availability, existing canine cancer registries are also affected by inherent limitations (Brønden et al. 2007). A first limitation is the presence of different and often incompatible data collection methods linked, for instance, to the absence of a universal system for canine cancer classification (Brønden et al. 2007; Nødtvedt et al. 2011). To overcome this issue, several canine cancer registries – among them the CANR, the Tulsa Registry, and the SCCR – have adopted a classification method based on the International Classification of Diseases for Oncology (ICD-O) in humans as well as for potential comparative studies involving human cancer (Grüntzig et al. 2015).

A second limitation is data completeness, which often leads to the systematic underestimation of canine cancer incidence (Brønden et al. 2007; Nødtvedt et al. 2011). Although gold standards for data collection are claimed to be a solution to the underreporting of canine cancer cases (Brønden et al. 2007), why there is underascertainment of canine cancer cases is, at present, relatively unclear. The reason is that canine cancer diagnostic examinations are not contingent on the dogs' but the owners' decision. Similar to humans, underascertainment of canine cancer cases can be linked to factors involving the use of healthcare (Fiscella et al. 2000). For example, the underascertainment of cancer cases relates to the affordability of veterinary healthcare (Bukowski and Wartenberg 1997). To tackle this issue, a number of regional canine cancer registries – for instance, the CANR (Dorn et al. 1968b) and the Animal Tumour Registry of the Venice and Vicenza provinces (Vascellari et al. 2009) – temporarily offered free diagnostic evaluations within their catchment area.

The last limitation of most existing canine cancer registries is the lack of ancillary information on the at-risk canine population (Nødtvedt et al. 2011). Except for the few countries where dog registration is a legal requirement, for example, in Denmark (Brønden et al. 2007) and Switzerland (Pospischil et al. 2013), at-risk populations are often estimated through telephone surveys (Dorn et al. 1968a) or information retrieved from pet insurance companies (Dobson et al. 2002). However, both estimation methods can be particularly unreliable when computing incidence rates involving small at-risk populations because minor errors in estimates can yield spurious rates (Brønden et al. 2007).

Some of the canine cancer data sources described previously have been employed in comparative studies of canine and human cancers. The main findings from these studies are discussed subsequently.

2.3.2. Comparative studies and main findings

Given the limitations of existing canine cancer registries, just a few comparative studies of canine and human cancers have adopted the methods of spatial epidemiology to date (Kimura et al. 2015). The most important contributions are from the 1960s and 1970s, when Dorn et al.

(1968a, b), Schneider et al. (1968), and Schneider (1970) developed a set of comparative studies based on the CANR. These authors studied the geographic distribution of the incidence of various canine cancers within selected counties of California, US, and correlated them with the geographic distribution of the same human cancers.

Thirty years later, O'Brien et al. (2000) carried out another groundbreaking comparative study on data from the Purdue Comparative Oncology Program. This study investigated the spatiotemporal distribution of four canine and human cancer types in Michigan, US, through a disease-clustering analysis, finding similar distributions. More recently, Pastor et al. (2009) explored the distribution of non-Hodgkin's lymphomas in dogs across French departments by successfully linking them to the presence of shared environmental exposures for human non-Hodgkin's lymphomas, waste incinerators, polluted sites, and sites storing radioactive waste.

The authors of these studies reported some of the limitations highlighted by Jacquez (2004). Among these, the most frequently mentioned are related to society and context as well as statistical inference (Dorn et al. 1968a; O'Brien et al. 2000; Pastor et al. 2009). In particular, O'Brien et al. (2000) lengthily discussed and requested further research on factors influencing the likelihoods of canine cancer diagnosis, such as, for instance, concerning geographic location, dog demographic characteristics, and typology of dog owners. Dorn et al. (1968a) also reported that uneven levels of underestimation of canine cancer incidence are a significant limitation in the generalizability of findings from canine to human cancers.

The same authors also described how most clinically-obvious canine cancer types are often more likely to be diagnosed compared to those affecting internal organs. Therefore, comparative studies involving some specific canine and human cancers are less suitable (Dorn et al. 1968b). Additional limitations associated, for instance, with data are not covered in these comparative studies mostly because they were performed with individual residential locations (Tedardi et al. 2015). Although spatial epidemiological studies comparing canine and human cancers are currently limited, purely epidemiological investigations are by far more common

(Schmidt 2009; Reif 2011). These studies mostly confirm common risk factors for the development of canine and human cancers, through case-control studies conducted with data collected within the histopathologic unit of selected veterinary hospitals (Bukowski and Wartenberg 1997).

For example, two early studies carried out by Reif and Cohen (1970) and Ragland and Gorham (1979) reported that dogs living in urban areas had a much higher prevalence of cancer of the tonsil than dogs living in rural areas. Reif et al. (1970) also observed a higher prevalence of lung diseases in dogs living in urban areas – making a case for similar harmful effects of urban air pollution on dogs and humans. Another study by Harbison and Godleski (1983) linked canine mesothelioma to the exposure to asbestos, and Glickman et al. (1983) demonstrated that this linkage was also related to the owner's exposure to asbestos in occupational or leisure settings. Further, Hayes et al. (1981) showed there to be a higher mortality linked to canine bladder cancer in dogs living near industrial activities.

Proximity to industrial activities was also linked to lymphoma by Gavazza et al. (2001), who also found an impact from the use of specific chemicals by dog owners. Adverse effects of chemicals on dogs' health were also evidenced by Glickman et al. (1989) where they observed that canine bladder cancer could be related to the use of pesticides. Furthermore, Hayes et al. (1981) showed that malignant lymphoma is linked to the use of herbicides. Reif et al. (1992) unsuccessfully tried to connect canine lung cancer to environmental tobacco smoke (ETS), but, in later studies, Reif et al. (1998) and Bukowski et al. (1998) made a connection between ETS and nasal cancer. This suggested that harmful impacts could affect an earlier part of the respiratory tract in dogs compared to humans. Reif et al. (1995) also linked an increased risk for lymphoma with dogs exposed to electromagnetic fields. Similar to the study carried out by Pastor et al. (2009), Marconato et al. (2009) observed there to be increased risk for cancer development in dogs living in proximity to hazardous waste disposal sites.

Given the variety of challenges and limitations highlighted by Jaquez (2004) and the literature review in the spatial epidemiology of canine cancer, the next section identifies relevant knowledge gaps.

2.4. KNOWLEDGE GAPS

Considering the elements highlighted in this literature review, the following specific knowledge gaps (KG) justify grounding the research questions presented in Section 1.1.1 into the four groups of challenges and limitations proposed by Jacquez (2004). In detail:

KG 1 *Society and context* play an essential role in the spatial epidemiology of canine cancer (Dorn et al. 1968b). The *implications of underascertainment of cancer cases* and the effects on the statistical performance of models of canine cancer incidence need to be better understood for effective environmental-sentinel applications.

KG 2 The spatial epidemiology of canine cancer is usually conducted at the individual level (O'Brien et al. 2000). For potential environmental-sentinel applications, it is necessary to better understand the *effects of spatial data aggregation* of canine cancer cases and potential explanatory variables on the statistical performance of models of canine cancer incidence.

KG 3 The spatial epidemiology of canine cancer usually considers small geographic scales (Tedardi et al. 2015). It is vital to underscore the *influences of spatial non-stationarity and geographic scale* on the statistical associations estimated in models of canine cancer incidence for potential environmental-sentinel applications.

KG 4 For *potential environmental-sentinel applications*, it is crucial to carefully contextualize statistical inference resulting from the spatial epidemiology of canine cancer (Scotch et al. 2009). This issue involves an in-depth understanding of the generalizability of statistical associations estimated in models of canine cancer incidence.

Owing to these crucial knowledge gaps, the next chapter, Chapter 3, presents the three cases studies with the SCCR data. These are developed to answer the proposed research questions.

3. THE CASE STUDIES

This chapter describes the case studies developed as part of this thesis in three distinct sections by detailing the materials, methods, results, and key findings. The first section addresses implications of underascertainment of cancer cases in the computation of canine cancer incidence. The second section investigates the effects of spatial data aggregation, namely the enumeration of canine cancer cases at the municipal level. The third section explores the influences of spatial non-stationarity and geographic scale.

3.1. IMPLICATIONS OF UNDERASCERTAINMENT OF CANCER CASES

This section addresses the first research question (RQ 1).

RQ 1 How do *society and context* challenge the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This is dealt with a case study published by Boo et al. (2017), which reports original research conducted by the author of this thesis. Gianluca Boo processed the data, developed and implemented the study design, interpreted the results, and wrote the first draft of the manuscript. Stefan Leyk edited the manuscript, contributed to the design, implementation, and interpretation of the results. Sara I. Fabrikant and Andreas Pospischil edited the manuscript and contributed to the interpretation of the results. Ramona Graf and Katrin Grüntzig collected and pre-processed the SCCR data. The content of the original manuscript is reported in a slightly altered form to better fit into the structure of this thesis.

3.1.1. Rationale

When assessing canine cancer incidence, the underascertainment of cancer cases may become paramount, because an indication of zero can originate

from the absence of diagnostic examinations within the sample unit. Such a zero is the manifestation of a structural phenomenon in the data source and should be discarded from any modeling effort (Hu et al. 2011; He et al. 2014). Still, structural zeros are challenging to identify and discard because they are often mistaken for sampling zeros. These are, in turn, resultant from diagnostic examinations performed within the sample units (Mohri and Roark 2005; Legendre and Legendre 2012). As a consequence of the persistent uncertainty surrounding the nature of zero incidence, little is known on the effects of structural zeros on models of canine cancer incidence to date.

To fill this critical knowledge gap, this case study evaluated the consequences of structural zeros on models of canine cancer incidence in Switzerland via a regression analysis framework. In doing so, we contrasted two modeling scenarios. The first scenario consisted of the complete enumeration of canine cancer incidence across all Swiss municipal units. The second scenario involved a filtered sample, which systematically discarded structural zeros — in other words, the municipal units where no diagnostic examination was performed during the year of interest.

This filtering step was allowed by the exceptionally rich attribution of the original canine cancer data source that contained information on the number of diagnostic examinations performed within each municipal unit in a specific year. By contrasting the statistical performance and predictive power of the two modeling scenarios in a cross-validation framework, new insights into the effects of structural zeros in models of canine cancer incidence are provided. These insights offer the ability to address a major challenge of the spatial epidemiology of canine cancer.

3.1.2. Materials and methods

Canine cancer diagnostic examinations and demographic indicators

The SCCR is a unique data source for the study of canine cancer comprising more than 120,000 diagnostic examinations performed in Switzerland between 1955 and 2008 (Grüntzig et al. 2015, 2016). This data source has been retrospectively assembled by the Collegium Helveticum Zurich for future comparative studies of canine and human cancers, and it is currently in

the process of being updated to include diagnostic examinations for the most recent years (Grüntzig et al. 2015, 2016).

As previous research has suggested that the accuracy and completeness of the SCCR data is reduced in earlier years (Grüntzig et al. 2015), we retrieved only the 7,057 diagnostic examinations performed in 2008. The diagnoses allowed for the ascertainment of 3,611 cancer cases. All types of malignant tumors were considered cancer cases, and dogs diagnosed with more than one cancer were considered single cases. For this case study, we enumerated both the number of diagnostic examinations and observed cancer cases at the municipal level based on the residential addresses stored in the diagnostic cases. This was done by linking the residential postcode to the unique identification number of the Swiss municipal units. This was done using a correspondence table available on the Swiss Federal Statistical Office website (SFSO 2017) as reference index to join these identifying attributes. The geocoding step enabled to successfully allocate more than 99.9% of the diagnostic cases to a municipal unit, while the remaining 0.1% were discarded.

To account for demographic risk factors within the at-risk canine population, we accessed demographic data on the 496,689 dogs living in Switzerland in 2008. The data was retrieved from the Swiss Canine Population Census, compiled by Animal Identity Service AG following the legal obligation of dog microchipping and registration established in Switzerland in 2006 (ANIS 2017). No exclusion criterion as to age or sex was adopted. For previous years, demographic data can be retrieved only for a limited number of municipalities, generally urban areas, or as estimates at the country level (Pospischil et al. 2013).

Based on the geocoded residential address of the registered dogs, we derived the size of the at-risk population (in number of individuals), the average age (in years), and the ratio of females per male dogs (in percent) within municipal units. The reason was that these independent variables are vital to the predisposition of several types of canine cancer (Eichelberg and Seine 1996; Lund et al. 1999; Michell 1999; Proschowsky et al. 2003).

Indicators of potential underascertainment of canine cancer cases

We assessed the urban character and socio-economic status of humans across Swiss municipalities because existing studies have inferred that these may be critical independent variables to account for confounding factors associated with potential underascertainment of canine cancer cases (Brønden et al. 2007; Bonnett and Egenvall 2010; Ponce et al. 2010). We first computed an indicator estimating human population densities at the municipal level (in 1,000 individuals per square kilometer) based on the extent of residential land within municipalities as the areal denominator. As such, we employed the Swiss Federal Statistical Office Census Data for 2008 (SFSO 2017) and information on the extent of the residential land derived from the building and dwelling survey conducted by the Swiss Federal Statistical Office in 2014 (SFSO 2017). Second, the socio-economic status was approximated based on average federal income tax information (in 1,000 Swiss Francs – CHF – per capita) collected by the Swiss Federal Tax Administration in 2008 (SFTA 2017).

We also derived an additional independent variable estimating the travel distance to veterinary care within municipalities (in kilometers) from a hectometric raster (i.e., with a 100m x 100m resolution) representing travel distance along roads (Delamater et al. 2012). In doing so, we assumed that increasing travel distance to veterinary services could be a crucial determinant for potential underascertainment of cancer cases. The raster was computed using the addresses of the 938 veterinary services registered in the official Swiss Yellow Pages online database in 2013 (Swisscom Ltd. 2017), and the Swiss road network in 2008 was derived from the VECTOR25 data model of the Swiss Federal Office of Topography (SFOT 2017). The distances to the closest veterinary service were averaged based on the location of their centroid to measure the average travel distance to veterinary care within a given unit (Bliss et al. 2012).

We considered more recent data on the addresses of veterinary services and the extent of residential land because data for 2008 was not easily accessible to us. Given the information provided by governmental agencies (FOPH 2017; SFSO 2017), this was seen as a reasonable compromise.

Table 1. Statistical distribution of the independent variables employed in the models across the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The table is adapted from Boo et al. (2017).

	(A) Complete enumeration				(B) Filtered sample			
Variable	**Median**	**IQR**	**Min**	**Max**	**Median**	**IQR**	**Min**	**Max**
Population Size (number of individuals)	118.0	181.5	1.0	146100	182.0	217.3	3.0	14610.0
Average Age (years)	6.7	0.9	3.0	13.0	6.6	0.8	4.9	9.6
Females per Male (percent)	50.9	7.4	0.0	100.0	50.7	6.2	20.0	75.9
Human Population Density (1,000 individuals per square kilometer)	1.2	1.1	0.0	15.8	1.5	1.3	0.0	15.7
Average Income Tax (1,000 CHF per capita)	0.6	0.6	0.0	25.6	0.7	0.6	0.0	15.3
Distance to Veterinary Care (kilometers)	3.0	2.9	0.4	33.0	2.7	2.4	0.4	30.0

Filtering out the structural zeros

Sampling is the process of selecting a representative set of individuals to make inferences about the entire population (Thompson 2012). In epidemiological research, this process involves selecting sampling units – defined as individuals or groups of individuals – to investigate relationships between a disease and its potential determinants, for instance, in cohort or case-control studies (Pearce 2012; Woodward 2013). In the first place, epidemiological studies need to be generalizable to inform about the disease determinants within the at-risk population from which the sampling units have been drawn (Pearce 2012; Woodward 2013). Various methods can be employed to define sampling units, typically using random and non-random designs (Cattin 1980; Banerjee and Chaudhury 2010).

While random sampling is designed to produce generalizable results, non-random sampling is critical because the representativeness for the entire at-risk population is not possible. Hence, the results of the epidemiological

study might not be generalizable (Cattin 1980; Banerjee and Chaudhury 2010). Although the sampling of enumerated data is rare in epidemiological research (Nejjari et al. 1993; Lawson 2006), we carried out a non-random selection of the municipal units where cancer diagnostic examinations were performed. This filtering step, discarding all structural zeros, is meant to draw a representative sample to evaluate the effects of underascertainment of cancer cases in models of canine cancer incidence (Cattin 1980; Banerjee and Chaudhury 2010). Therefore, in parallel, we also fit the model based on the complete enumeration of cancer cases across all Swiss municipal units.

We then compared the distributions utilizing descriptive statistics and assessed the statistical performance and predictive power of the two modeling scenarios. This comparison was meant to evaluate potential changes associated with our filtering step, and identify consequences of structural zeros on the model of canine cancer incidence.

Modeling canine cancer incidence

We fit the observed canine cancer incidence within a Poisson regression framework as this is one of the most common models for assessing the incidence of rare diseases, such as cancer (Frome 1983; Frome and Checkoway 1985).

However, the incidence data may deviate from a standard Poisson distribution, thus introducing uncertainty via the coefficient estimation (Cameron and Trivedi 1990; Berk and MacDonald 2008). Still, we decided against testing alternative regression frameworks, such as negative binomial (Hardin and Hilbe 2007; Berk and MacDonald 2008), zero-inflated and hurdle (Hu et al. 2011; He et al. 2014) models. This was because the coefficients accommodating different statistical distributions impede a direct comparison between the two modeling scenarios (Preisser et al. 2012; Arab 2015). Moreover, the relatively simple structure of the Poisson regression framework enables a more straightforward assessment of changes in the coefficient estimates of each independent variable (Arab 2015).

Given these preliminary considerations, we fit the observed canine cancer incidence (y) through the following independent variables (x) – *Population Size* (in number of individuals), *Average Age* (in years), *Females*

per Male (in percent), *Average Income Tax* (in 1,000 CHF per capita), *Human Population Density* (in 1,000 individuals per square kilometer), and *Distance to Veterinary Care* (in kilometers). The fit canine cancer incidence ($\hat{y}$) were log-transformed according to **Equation 1**. In this equation, α is the intercept, β the multiplicative coefficient estimated for each independent variable, and ε the error term (Frome 1983; Frome and Checkoway 1985).

$$\log(\hat{y}(y|x)) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \quad \text{(Equation 1)}$$

To contrast the two modeling scenarios, we evaluated significance levels (α=.05) and changes in the multiplicative effects (i.e., exp(β), in percent) associated with the different coefficient estimates as well as the associated proportion of variance reduction (η^2) (Levine and Hullett 2002). In doing so, we focused on potential changes occurring between two sets of independent variables – (1) *Population Size*, *Average Age* and *Females per Male*, and (2) *Average Income Tax*, *Human Population Density*, and *Distance to Veterinary Care*. This is because these are expected to inform about two distinct processes – demographic risk factors for canine cancer and confounding factors associated with potential underascertainment of cancer cases.

Next, we computed the McFadden pseudo-R-squared ($R^2_{McFadden}$) as a measure of the statistical performance of the two modeling scenarios (McFadden 1973). A $R^2_{McFadden}$ statistic approaching 0 indicates decreasing performance, while a value of 1 indicates perfect performance (Cameron and Windmeijer 1996). We also mapped the residuals resulting from the two modeling scenarios to identify municipal units of poor model prediction along with potential spatial non-stationarity with regards to the statistical associations (Fotheringham et al. 1996; Brunsdon et al. 1996). We opted to examine Pearson residuals because they can highlight a critical lack of model fit, namely when the absolute values exceeded 2.0, and especially 3.0 (Cameron and Windmeijer 1997).

Validating the modeling scenarios

Given that the cancer incidence fit with the two modeling scenarios was expected to present different distributions, we employed a cross-validation method based on 1,000 model iterations to contrast statistical performance

and predictive power (Snee 1977; Picard and Cook 1984). Cross-validation enables enabled assessment of how well a selected model will generalize to a different dataset – a critical issue for potential comparative studies of canine and human cancers (St. Sauver et al. 2012; Kukull and Ganguli 2012). For each model iteration, we randomly fit 80% of the municipal units (i.e., the training set) to predict the remaining 20% (i.e., the validation set) (Snee 1977; Picard and Cook 1984).

We then assessed central tendency and spread of the multiplicative effects associated with the different coefficient estimates across iterations using boxplots (Williamson et al. 1989). The goal with this was to evaluate the stability of statistical associations across iterations, and therefore statistical performance (Snee 1977; Picard and Cook 1984). We also computed measures of predictive power by averaging the mean absolute error (MAE) and the root mean square error (RMSE) across iterations (Willmott 1981; Hyndman and Koehler 2006). As shown in **Equation 2**, the MAE is an absolute measure of the error, defined as the difference between the predicted ($\hat{y}$) and observed (y) canine cancer incidence (Willmott 1981).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \quad \textbf{(Equation 2)}$$

To better understand the distribution of the error across iterations, we also computed the 50th, 90th, and 95th percentiles of the MAE, which were also averaged for the two modeling scenarios (Willmott 1981). We then computed the RMSE, which, as presented in **Equation 3**, is defined as the square root of the squared error (Willmott 1981).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad \textbf{(Equation 3)}$$

In computing the average RMSE, we discarded the iterations resulting in RMSE outliers through a standard single-step outlier-detection procedure (Hawkins 1980) as this measure is known to be sensitive to large errors (Chai and Draxler 2014). We finally assessed potential correlations between RMSE and the independent variables utilized to fit the training sets for the two modeling scenarios to assess whether the latter may drive variations in the error measures. This exploratory step was performed through Spearman's ρ-

correlation tests (Spearman 1904), where significant relationships (α=.05) were further investigated with scatterplots.

3.1.3. Results

Defining the modeling scenarios

Figure 3 depicts the geographic distribution of observed canine cancer incidence in Switzerland in 2008 fit with the two modeling scenarios described earlier. **Figure 3A** defines the incidence for the modeling scenario based on the complete enumeration (i.e., including all Swiss municipalities) where 1,298 municipal units out of 2,350 indicate zero incidence. **Figure 3B** portrays the filtered incidence data, where 939 municipal units exhibited structural zeros because no diagnostic examinations were performed. These municipal units are labeled as "no data." As a consequence of the filtering step, only 359 municipal units exhibited zero incidence. These numbers demonstrate that structural zeros are widespread in the data retrieved from the SCCR in 2008.

As such, the statistical distributions could be substantially different between the two modeling scenarios. In fact, the mean and median of the incidence data were 1.5 and 0.0 for the complete enumeration compared to 2.5 and 1.0 for the filtered sample. The coefficient of variation of the incidence data was 334% for the complete enumeration and 250% for the filtered sample. These measures show that, for the two modeling scenarios, the incidence deviated from a standard Poisson distribution, and this deviation was greater in the complete enumeration because of the higher coefficient of variation. **Figure 3B** also suggests that zeros were not randomly distributed across the study area. The reason was that the regions of municipal units with sampling zeros (i.e., zero incidence) and structural zeros (i.e., "no data") could be identified in rural regions, such as the Alps (South) and Jura Mountains (North-West).

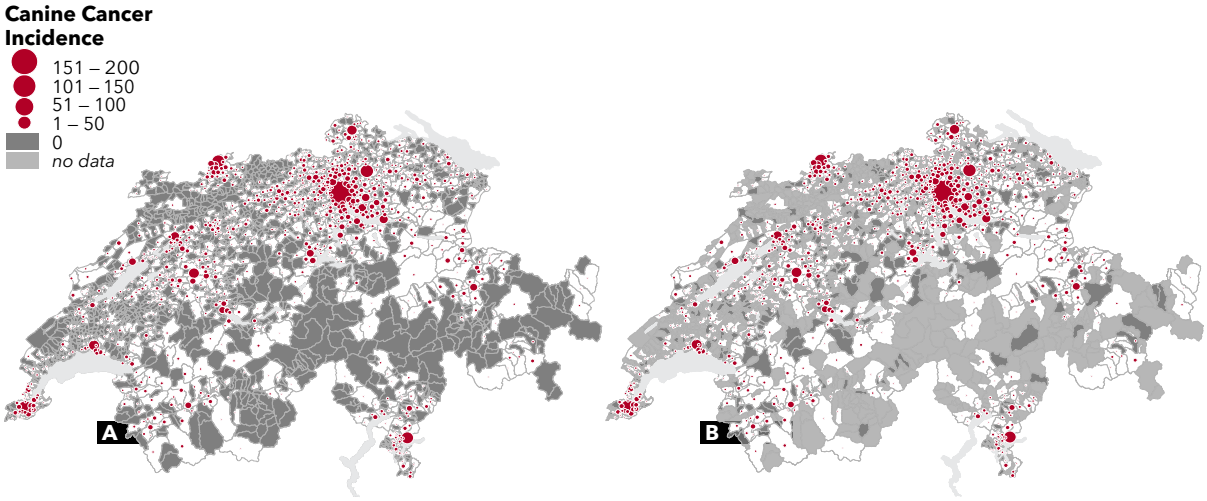

Figure 3. Geographic distribution of the canine cancer incidence in Switzerland in 2008 across the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The data is classified according to the quantile classification method. The figure is adapted from Boo et al. (2017).

Table 2 outlines that, when fitting the canine cancer incidence in a Poisson regression framework, all coefficient estimates were statistically significant ($P<.05$) and remained very similar across the two modeling scenarios.

The coefficient estimates suggest that *Average Age* involved negative relationships, namely that for each increasing year of age, the incidence decreased by 27.4% (complete enumeration) and 28.1% (filtered sample). Conversely, *Population Size* and *Females per Male* both produced positive relationships – for each extra individual and percentage unit of females, the incidence rose by 25.9% (complete enumeration) and 27.1% (filtered sample) as well as 1.0% (complete enumeration) and 2.0% (filtered sample), respectively.

Average Income Tax and *Human Population Density* both had positive relationships, implying that for each 1,000 CHF per capita and 1,000 individuals per square kilometer, the incidence increased by 11.6% (complete enumeration) and 12.6% (filtered sample) as well as 25.9% (complete enumeration) and 20.9% (filtered sample), respectively. Lastly, *Distance to Veterinary Care* exhibited negative relationships, indicating that for each kilometer of distance, the incidence decreased by 18.1% (complete enumeration) and 10.4% (filtered sample).

Table 2 also shows that, altogether, the independent variables accounting for confounding factors associated with potential underascertainment of cancer cases had a lower proportion of variance reduction for the filtered sample (η^2=0.36) compared to the complete enumeration (η^2=0.46). When assessing the statistical performance of the two modeling scenarios, the $R^2_{McFadden}$ measures indicated a slightly increased statistical performance of the model based on the filtered sample ($R^2_{McFadden}$=0.32) compared to the complete enumeration ($R^2_{McFadden}$=0.31).

Table 2. Coefficient estimates, P-values, and proportions of variance reduction (η^2) across the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The table is adapted from Boo et al. (2017).

	(A) Complete enumeration			**(B) Filtered sample**		
Coefficient	**Estimate**	***P***	**η^2**	**Estimate**	***P***	**η^2**
Population Size	0.23	<.001	0.54	0.24	<.001	0.62
Average Age	–0.32	<.001	0.00	–0.33	<.001	0.01
Females per Male	0.01	.03	0.00	0.02	<.001	0.01
Human Income Tax	0.11	<.001	0.07	0.12	<.001	0.08
Human Population Density	0.23	<.001	0.28	0.19	<.001	0.23
Distance to Veterinary Care	–0.20	<.001	0.11	–0.11	<.001	0.05

Figure 4 depicts the geographic distribution of the Pearson residuals for the two modeling scenarios. **Figure 4A** shows that, for the complete enumeration, most municipal units located in the Alps (South) and Jura Mountains (North-West) were characterized by acceptable model over-estimations as the residuals spanned between –1.9 and 0.1. **Figure 4B** shows an increased predictive power for the filtered sample. This was because most municipal units with residuals above 2.0 and below –2.0 in the complete enumeration had residuals between –1.9 and 1.9 in the filtered sample. The two modeling scenarios also exhibited several regions with residuals above 3.0 located within urban agglomerations of Zurich, Basle, and Berne.

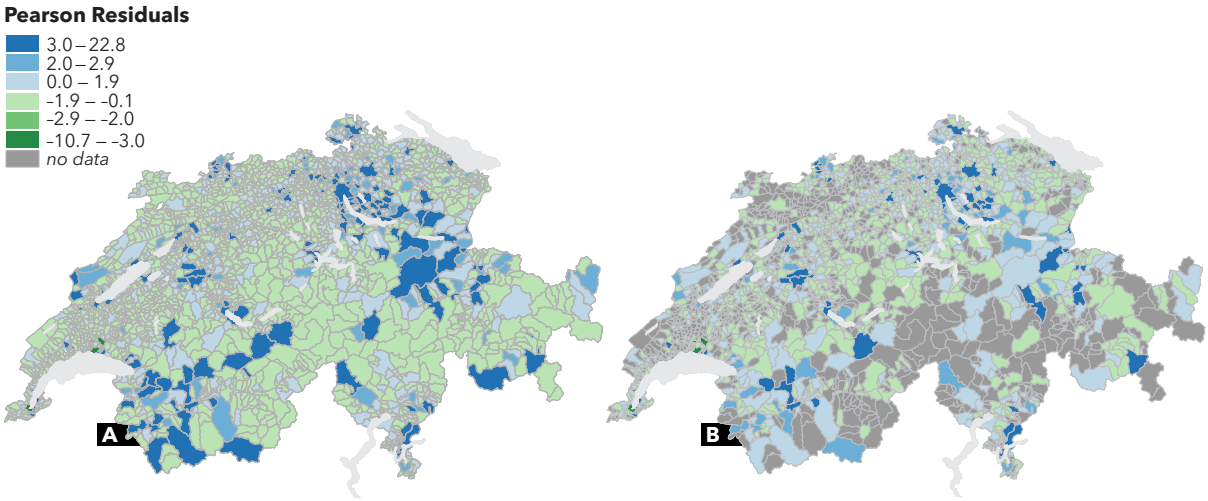

Figure 4. Geographic distribution of the Pearson residuals across the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The data is classified according to the fixed classes classification method. The figure is adapted from Boo et al. (2017).

Comparing the modeling scenarios

We further contrasted the statistical performance and predictive power of the two modeling scenarios through model cross-validation. This was based on a training-/validation-set ratio of 1881/470 municipal units for the complete enumeration and 1130/282 municipal units for the filtered sample.

The boxplots in **Figure 5** depict the spread of the multiplicative effects associated with the different coefficient estimates over the 1,000 model iterations for the two modeling scenarios. The median is shown as a thick horizontal line, the interquartile range is indicated as a grey box and the minimum and maximum multiplicative effects are delimited by the whiskers. In **Figure 5**, the median values for the two modeling scenarios looked very similar to the multiplicative effects associated with the coefficient estimates presented in **Table 2**, therefore indicating overall stability across iterations.

However, when comparing the distributions across the two modeling scenarios, *Population Size*, *Average Income Tax*, *Human Population Density* and *Distance to Veterinary Care* showed a decreased spread for the filtered sample. This indicates an enhanced stability of multiplicative effects across iterations for the independent variables accounting for the size of the different at-risk populations and the confounding factors associated with

underascertainment of cancer cases, thus suggesting an improved statistical performance for the filtered sample.

Lastly, demographic variables such as *Average Age* and *Females per Male* featured similar spreads for the two modeling scenarios, possibly because the large portion of zeros in the complete enumeration stabilized the coefficient estimates toward zero. Despite the similar spread, *Females per Male* showed both positive and negative multiplicative effects for the complete enumeration, thus denoting contrasting relationships.

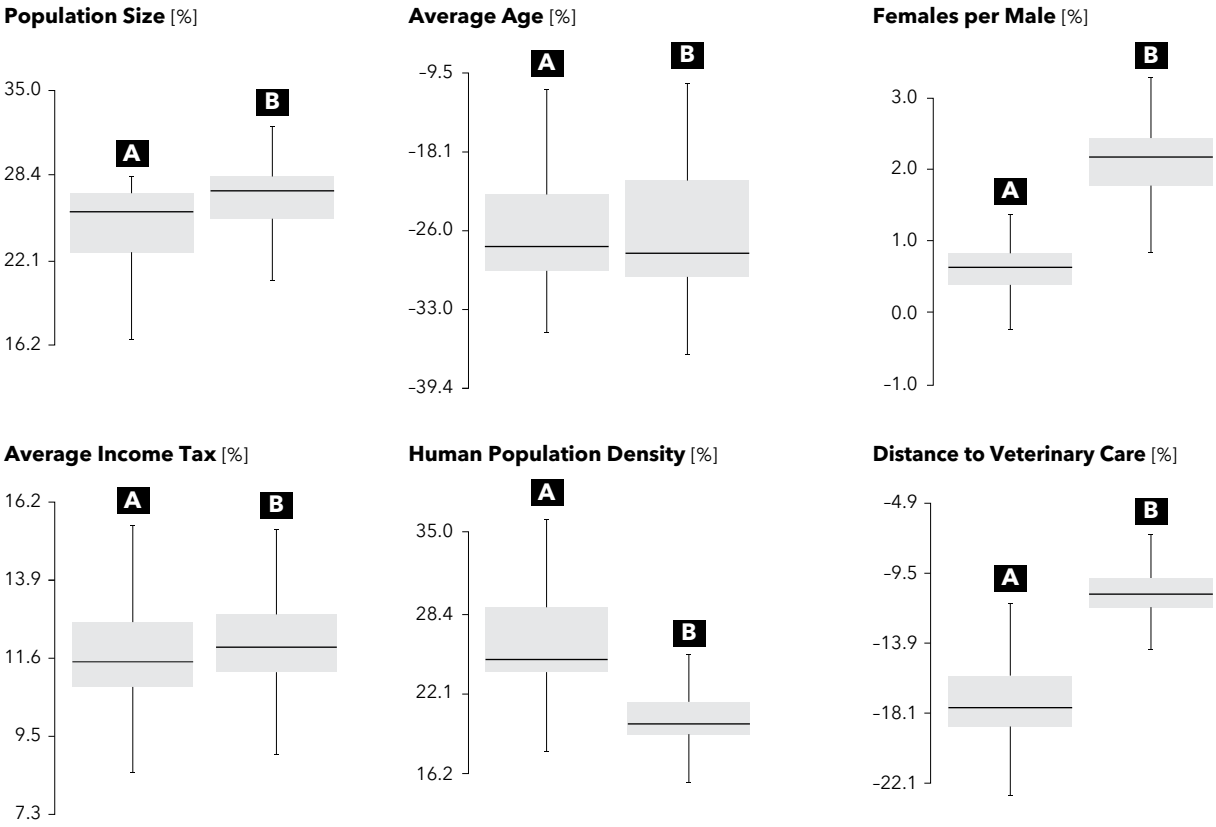

Figure 5. Distribution of the multiplicative effects associated with the different coefficient estimates across the iterations through the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The figure is adapted from Boo et al. (2017).

Table 3 shows that the average MAE for the complete enumeration was four times larger than for the filtered sample. The averaged percentiles of the MAE show that the average error distribution was heavily skewed in the complete enumeration as, on average, only 5% of the errors (i.e., above the 95th percentile) were accountable for a higher MAE. In the filtered sample, the error distribution appeared less skewed because of a general decrease of the

error magnitudes. Furthermore, this skewed error distribution also affected the calculation of the average RMSE for the complete enumeration.

The outlier detection procedure revealed six iterations with RMSE between 3,700 and 159,300 in the complete enumeration, and three iterations with RMSE between 5,700 and 9,000 in the filtered sample. These RMSE outliers were at least 10 times higher than the average RMSE for the two modeling scenarios and were removed. After discarding these outliers, the average RMSEs became more meaningful for comparing the two modeling scenarios and still confirmed a higher predictive power for the filtered sample.

Table 3. MAE and RMSE measures averaged across iterations through two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The table is adapted from Boo et al. (2017).

	Average MAE				**Average RMSE**	
Modeling scenario	**Raw**	**50th**	**90th**	**95th**	**Raw**	**No outliers**
(A) Complete enumeration	19.35	0.90	2.58	4.29	362.15	20.25
(B) Filtered sample	4.29	1.26	3.32	4.98	40.88	17.28

Lastly, we computed Spearman's ρ correlations across the model iterations between the RMSE and the independent variables utilized to fit the training sets. This was to better comprehend the role of different independent variables as drivers of variation in the RMSE measures. The only significant correlation with the RMSE was found for the averaged *Population Size*, with a ρ of –0.72 ($P<.001$) for the complete enumeration and a ρ of –0.76 ($P<.001$) for the filtered sample. Such strong negative correlation between the averaged *Population Size* employed in the training sets and the RMSEs for the two modeling scenarios is presented in **Figure 6**.

In **Figure 6**, the trend of the correlation is highlighted by the grey surface representing the conditional mean smoothed through a linear model fit. Again, the trend surfaces demonstrated that smaller average population sizes in the training sets result in higher RMSE, and thus in lower predictive power. For both modeling scenarios, the RMSE seemed to be partly independent of the average population sizes utilized in the training set, as there are four distinct RMSE groups centered around 0.0, 5.0, 50.0, and 150.0.

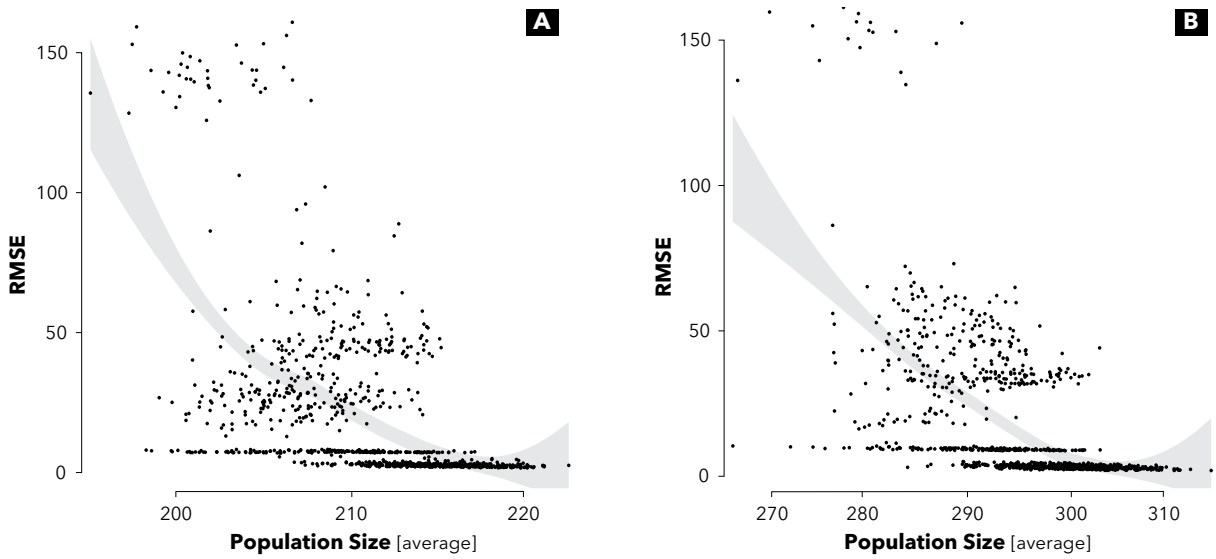

Figure 6. Average population size versus RMSE across the iterations through the two modeling scenarios – (**A**) complete enumeration and (**B**) filtered sample. The grey surface represents the trend in the data based on the conditional mean smoothed through a linear model fit. The figure is adapted from Boo et al. (2017).

3.1.4. Summary and key findings

This case study examines potential causes of underascertainment of cancer cases and relevant effects on the statistical performance and predictive power of models of canine cancer incidence. For this purpose, two scenarios for modeling incidence data retrieved from the SCCR were defined. The first scenario was based on the complete enumeration of incidence data for all Swiss municipal units. The second scenario was based on a filtered sample that systematically discarded structural zeros in those municipal units where no diagnostic examination had been performed. Using cross-validation, we then evaluated and contrasted the statistical performance and predictive power of the two modeling scenarios. This comparative assessment revealed the following:

- Structural zeros impacted the statistical distribution of the data. These were mostly located in proximity to sampling zeros – in the Alps and Jura mountains;
- In spite of different distributions, the two modeling scenarios did not exhibit substantial variations in the statistical performance (i.e., $R^2_{McFadden}$) and the coefficient estimates;

- Model cross-validation enables detecting the increased instability of the coefficient estimates and larger measures of error (i.e., MAE and RMSE) in the complete enumeration; and
- In both modeling scenarios, smaller error measures were highly correlated with the increasing size of the sample employed for training the model.

The results of this case study will be interpreted and discussed in Section 4.1.1. The next section features the second case study of this thesis, which deals with the challenges and limitations associated with spatial data aggregation.

3.2. EFFECTS OF SPATIAL DATA AGGREGATION

This section addresses the second research question (RQ 2).

RQ 2 What are the implications of using *spatial data* in the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This is performed through a case study submitted by Boo et al. (2018b), which reports original research conducted by the author of this thesis. Gianluca Boo processed the data, developed and implemented the study design, interpreted the results, and wrote the first draft of the manuscript. Stefan Leyk edited the manuscript, contributed to the design, implementation, and interpretation of the results. Sara I. Fabrikant and Andreas Pospischil edited the manuscript and contributed to the interpretation of the results. Ramona Graf and Katrin Grüntzig collected and pre-processed the SCCR data. The content of the original manuscript is reported in a slightly altered form to better fit into the structure of this thesis.

3.2.1. Rationale

This case study further investigated potential underascertainment of canine cancer cases in a regression analysis framework by exploring statistical associations between canine cancer incidence rates and selected independent variables. However, this modeling effort is likely to be affected by the effects of spatial data aggregation, such as the MAUP because the incidence is computed by enumerating canine cancer cases within municipal units (Openshaw 1984; Cressie 1996). The MAUP affects statistical analysis when using spatially aggregated data because summary statistics change according to the shape and spatial extent of the enumeration unit (Openshaw 1984).

Spatial data aggregation further implies that the resulting summary statistics are homogeneously distributed within the enumeration units and that sharp changes occur across their boundaries (Wright 1936; Eicher and Brewer 2001). This assumption is unrealistic when modeling canine cancer incidence because, similar to humans, populations and diseases are not

randomly distributed across space (Cleek 1979). To evaluate the impacts of spatial data aggregation, and, simultaneously, explore uncertainty in the original canine cancer data source, we contrasted models based on two enumeration types.

The first type were municipal units while the second were dasymetrically refined units, defined as the portion of residential land within municipalities. This analytical framework was meant to explore uncertainty in the canine cancer registry data while examining whether statistical associations and statistical performance are affected by dasymetric refinement. Given that effects of spatial data aggregation involving canine cancer cases are mostly unknown, to date, this was considered an important stepping stone toward a better understanding of this source of uncertainty in the spatial epidemiology of canine cancer.

3.2.2. Materials and methods

Canine cancer diagnostic examinations and demographic indicators

The SCCR currently stores canine cancer diagnostic examinations from 1955 to 2008 for the entire country. The data was retrospectively assembled by the Collegium Helveticum Zurich and, and as described earlier on, is currently being updated (Grüntzig et al. 2015, 2016). As of this moment, the SCCR consists of 121,936 canine cancer diagnostic examinations from post-mortem and biopsy samples, performed at the Vetsuisse institutes of veterinary pathology in Berne and Zurich, and at a private diagnostic laboratory located in the Zurich area (Grüntzig et al. 2015). This case study is based on the 3,611 canine cancer diagnostic cases ascertained during the year 2008. These were enumerated at the municipal level, using the residential addresses stored in the diagnostic data, by linking the residential postcode to the unique identification number of the Swiss municipal units (SFSO 2017). All types of malignant tumors were considered as cancer cases, and dogs diagnosed with more than one cancer were considered single cases.

To yield demographic indicators based on the at-risk canine population in Switzerland for 2008, we also accessed the Swiss Canine Population

Census. This was compiled by Animal Identity Service AG (ANIS 2017) following the Swiss legal obligation of dog microchipping and registration established in 2006 in Switzerland (Pospischil et al. 2013). No exclusion criterion in terms of age or sex was adopted. Based on the residential address of the registered dogs, we computed demographic indicators describing the size of the at-risk population (in number of individuals), the average age (in years), the ratio of females per male dogs (in percent), and the ratio of mixed breed dogs (in percent) at the municipal level. Mixed breed dogs were defined according to the standards of the Fédération Cynologique Internationale (FCI) (FCI 2017).

While age and sex have similar relationships to cancers in dogs and humans (Owen 1979; Pinho et al. 2012), different cancer incidences among dog breeds could be a potential confounding factor for future comparative studies of dog and human cancers (Michell 1999; Proschowsky et al. 2003).

Indicators of potential underascertainment of canine cancer cases

We employed indicators of urban character and socio-economic status to estimate confounding factors associated with a different use of veterinary care, and, thus, potential underascertainment of canine cancer cases (Brønden et al. 2007; Nødtvedt et al. 2011; O'Neill et al. 2014).

First, we assessed the urban character of municipalities, because the use of veterinary care was expected to be more frequent in urban locations. This was calculated based on human population density (in 1,000 people per square kilometer) with population census data at the municipal level for the year 2008 (SFSO 2017). Second, we assumed that municipalities characterized by higher socioeconomic status are more likely to possess financial means for regular veterinary check-ups potentially resulting in cancer diagnosis. We considered a surrogate to describe the socio-economic status of municipalities through average federal income tax information for 2008 (in 1,000 CHF per capita) (SFTA 2017).

At last, we derived a measure of distance to veterinary care (in kilometers) as we expected that greater road distance to veterinary services would result in increased underascertainment of cancer cases. This independent variable was based on the addresses of the 938 veterinary

services active in 2013 (Swisscom Ltd. 2017). We created a hectometric distance-grid (i.e., with a 100m x 100m resolution) representing distances along roads (Delamater et al. 2012) using the Swiss road network in 2008, which had been extracted from the VECTOR25 data model of the Swiss Federal Office of Topography (SFOT 2017). Municipal-level average road distances to the closest veterinary service were computed by averaging the distance-grid values based on the location of their centroid (Bliss et al. 2012).

We utilized the address locations of registered veterinary services in 2013 because data for 2008 were not readily available to us. However, information issued by the Swiss Registry of Medical Professions confirmed negligible changes in the number of licensed veterinarians over this period (FOPH 2017).

The different types of enumeration unit

We investigated uncertainty in the SCCR data at the municipal level because this is the finest administrative level. We utilized the boundaries of the 2,350 Swiss municipal units as derived from the swissBOUNDARIES3D vector data model of the Swiss Federal Office of Topography (SFOT 2017). We retrieved municipal unit boundaries for 2014 because the SCCR data for all prior years have been systematically encoded and allocated to the municipal units existing in 2014. To evaluate possible improvements in statistical performance considering the effects of spatial data aggregation, we performed a dasymetric refinement of the municipal units based on the spatial extent of residential land within each municipality (Eicher and Brewer 2001).

The use of residential land as an ancillary variable for dasymetric refinement is based on the assumption that dogs and humans share the same living environment (Reif 2011). We derived residential land data from the building and dwelling survey conducted by the Swiss Federal Statistical Office in 2014. The data is available as a hectometric grid (i.e., with a 100m x 100m resolution). Grid cells are classified as residential land if they intersect the centroid of at least one residential building. The survey retrieves information on characteristics and geographic coordinates of the buildings from the Federal Register of Buildings and Dwellings (RBD) (SFSO 2017).

We made use of more recent information on residential land because data for 2008 were not available. However, differences between the

corresponding years were reported to be minimal because of the increasing densification of residential land parcels, especially in urban areas (SFSO 2017). Using the building and dwelling survey data for 2014 was an acceptable compromise.

Table 4. Statistical distribution of the independent variables employed in the models based on the two enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. The table is adapted from Boo et al. (2018b).

	(A) Municipal units				**(B) Dasymetrically refined units**			
Variable	**Median**	**IQR**	**Min**	**Max**	**Median**	**IQR**	**Min**	**Max**
Average Age (years)	6.7	0.9	3.0	13.0	6.7	0.9	3.0	13.0
Females per Male (percent)	50.9	7.4	0.0	100.0	50.9	7.4	0.0	100.0
Mixed-Breed Ratio (percent)	27.4	10.1	0.0	100.0	27.4	10.1	0.0	100.0
Average Income Tax (1,000 CHF per capita)	1.2	1.1	0.0	15.8	1.2	1.1	0.0	15.8
Human Population Density (1,000 individuals per square kilometer)	0.2	0.3	0.0	11.3	0.6	0.6	0.0	25.6
Distance to Veterinary Care (kilometers)	3.0	2.9	0.4	33.0	2.7	2.8	0.3	32.2

Dasymetric refinement of the enumeration units

We employed a dasymetric framework to evaluate possible improvements in our regression analysis framework, which could be linked to reduced effects associated with spatial data aggregation and subsequent recomputation of the explanatory variables associated with certain risk factors. Dasymetric refinement is a cartographic method meant to elicit more accurate geographic distributions of enumerated data considering geographic context (Wright 1936; Eicher and Brewer 2001). It is performed using ancillary spatial variables that are assumed to be related to the outcome and expected to reflect the geographic distribution of the data more accurately.

Figure 7 illustrates a framework for dasymetric refinement of population data within administrative units using residential land as a limiting ancillary variable (**Figure 7B**) – a dasymetric mapping technique referred to as the binary method (Mennis 2009). Compared to the choropleth map based on administrative units (**Figure 7A**), the binary dasymetric map (**Figure 7C**) produces more accurate geographic distributions of the population data, hence yielding more robust density estimates (Eicher and Brewer 2001). Such a dasymetric refinement is constrained by the pycnophylactic property, in other words, population counts of dasymetrically refined units should maintain the same values of the original enumeration units (Tobler 1979).

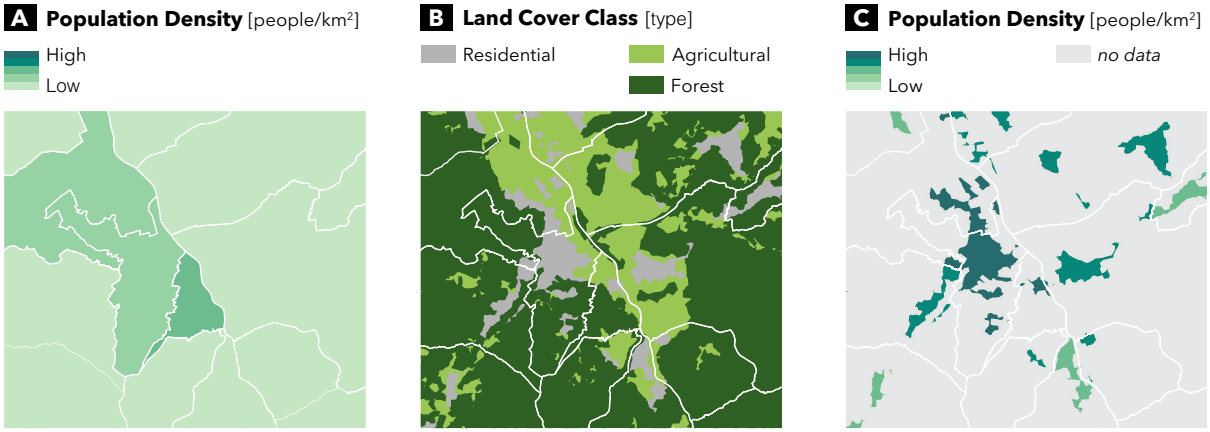

Figure 7. Example of binary dasymetric refinement of population data within residential land – (**A**) population density computed within administrative units is refined based on (**B**) the location of residential land to recompute (**C**) population density within dasymetrically refined units. The figure is adapted from Boo et al. (2018b).

To date, various types of data have been tested as ancillary variables for dasymetric refinement, including land use (Mennis 2003; Mennis and Hultgren 2006; Leyk et al. 2013), road density (Reibel and Bufalino 2005), remote-sensing imagery (Zandbergen 2011), parcel data (Tapp 2010; Nagle et al. 2014; Zoraghein et al. 2016), address points (Zandbergen 2011), and dwelling survey data (Boo et al. 2015). In this case study, we refined the enumeration units in a binary fashion to their portion of residential land – similar to the example in **Figure 7**.

The cells representing residential land were allocated to municipal units according to the location of their cell centroid. Once allocated to a municipality, the cells were dissolved and the resulting spatial extents were

employed as dasymetrically refined enumeration units. These units were utilized to enumerate the canine cancer incidence and the indicators implemented as independent variables within the regression modeling framework. The only two differences between regression models based on municipal units and those based on dasymetrically refined units were in the independent variables involving recomputed density and distance indicators.

This is because only these independent variables change according to the modified spatial extent and relative location of the enumeration unit (Openshaw 1984; Cressie 1996). While the calculation of density indicators is a natural application of dasymetric mapping to reduce effects of the MAUP (Eicher and Brewer 2001; Mennis 2009), computing distance indicators based on dasymetrically refined units involves a typical change of support (i.e., downscaling) that is not subject to the pycnophylactic propriety (Cressie 1993, 1996). The impact of these recomputations on statistical associations and performance is central in the regression modeling framework, therefore informing about potential improvements that could be related to spatial data disaggregation within more meaningful enumeration units.

Modeling canine cancer incidence rates

We employed a Poisson regression framework as a baseline model to fit canine cancer incidence rates. To adjust the observed canine cancer incidence (y) for the underlying at-risk canine population, the variable *Dog Population* (in number of individuals) was employed as an offset (e) – a constant of proportionality for computing incidence rates (Frome 1983; Frome and Checkoway 1985).

Canine cancer incidence rates were fit with the following independent variables (x): *Average Age* (in years), *Females per Male* (in percent), *Mixed-Breed Ratio* (in percent), *Average Income Tax* (in 1,000 CHF per capita), *Human Population Density* (in 1,000 individuals per square kilometer), and *Distance to Veterinary Care* (in kilometers). The fit canine cancer incidence rates ($\hat{y}$) were log-transformed according to **Equation 4**. In the equation, α is the intercept, β the multiplicative coefficient estimated for each independent variable, and ε the error term (Frome 1983; Frome and Checkoway 1985).

$$\log(\hat{y}(y|x)) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \log(e) + \varepsilon \quad \textbf{(Equation 4)}$$

We investigated significance levels (α=.05) and changes in the multiplicative effects (i.e., exp(ß) in percent) associated with the different coefficient estimates as well as the associated percentage of variance reduction (η^2) (Levine and Hullett 2002). On the one hand, this was meant to explore the statistical associations to the confounding factors associated with potential underascertainment of cancer cases. On the other hand, as this phenomenon can also introduce uncertainty in the multiplicative effects of the demographic variables, we were also able to detect potential mismatches with prior findings on demographic risk factors for canine cancer incidence (Bronson 1982; Eichelberg and Seine 1996; Michell 1999; Proschowsky et al. 2003).

Considering that our baseline Poisson model is based on the highly restrictive assumption of equidispersion, in other words, with variance equal to the mean of the data (Cameron and Trivedi 1990), we compared four different regression frameworks for count data – (**1**) Poisson model, (**2**) Poisson model with zero-inflation extension, (**3**) negative binomial model, and (**4**) negative binomial model with zero-inflation extension (Zeileis et al. 2008). While the negative binomial models (i.e., **3** and **4**) relax the assumption of equidispersion by accounting for a variance greater than the mean (i.e., overdispersion) (Cameron and Trivedi 1990), the zero-inflation extensions (i.e., **2** and **4**) place excess zeros in a separate logistic regression model with a binary outcome (i.e., zero versus non-zero counts) (Zeileis et al. 2008).

Modeling these excess zeros separately has the important advantage of providing potential insights into the nature of structural zeros (Lambert 1992; Zeileis et al. 2008). To avoid model overspecification, we first implemented all the independent variables presented before in the zero-inflation extensions but finally retained only the significant (α=.05) ones.

We assessed the statistical performance of the regression models through the AIC (Akaike 1974). This was through a systematic pairwise assessment of the relative likelihood, that is, the probability that a model minimizes the estimated information loss similar to the model with the lower AIC (Burnham and Anderson 2003). To assess the significance (α=.05) of

improvement of one model over another, we also performed systematic pairwise likelihood-ratio tests (Lewis et al. 2011). This form of comparison was meant to overcome the use of the Vuong test (Vuong 1989) as several concerns about its validity have been raised (Lewis et al. 2011; Wilson 2015).

Finally, we determined changes in the statistical associations between the models based on the two enumeration types. This implied a comparison of the multiplicative effects and size (η^2) of significant (α=.05) coefficient estimates (Zeileis et al. 2008). In so doing, we focused, in particular, on *Human Population Density* and *Distance to Veterinary Care* because both were expected to be affected by the effects of spatial data aggregation.

3.2.3. Results

Assessing the effects of dasymetric refinement

Figure 8 shows the portion of residential land within municipal units reflecting changes in the spatial extent of enumeration units associated with binary dasymetric refinement. These changes impact the recomputation of density indicators, namely *Human Population Density*. As expected, the greatest differences in the spatial extent occurred in the Alps (South) and Jura Mountains (North-West) which had very small residential land proportions, mostly less than 10.0%. In contrast, higher residential land proportions, between 10.0% and 59.9%, generally characterize the Central Plateau, with peak proportions exceeding 60.0% for the larger urban agglomerations like Zurich, Geneva, or Basle.

Figure 8 also portrays changes in the relative location of the enumeration unit centroids associated with the change of support resulting from dasymetric refinement. The width and direction of the purple arrows symbolize the magnitude and direction of displacement of the centroids from the municipal units (base of the arrow) to the dasymetrically refined units (point of the arrow). These changes in relative location impact the recomputation of distance indicators, specifically *Distance to Veterinary Care*. Again, as expected, the greatest centroid shifts, between 2.5 and 15.5 km, occurred in the Alps, while centroid shifts in the Central Plateau are much smaller, between 0.0 and 2.4 km. This can be explained by the small spatial

extent of the municipal units and more homogeneous spread of settlements within the municipal units.

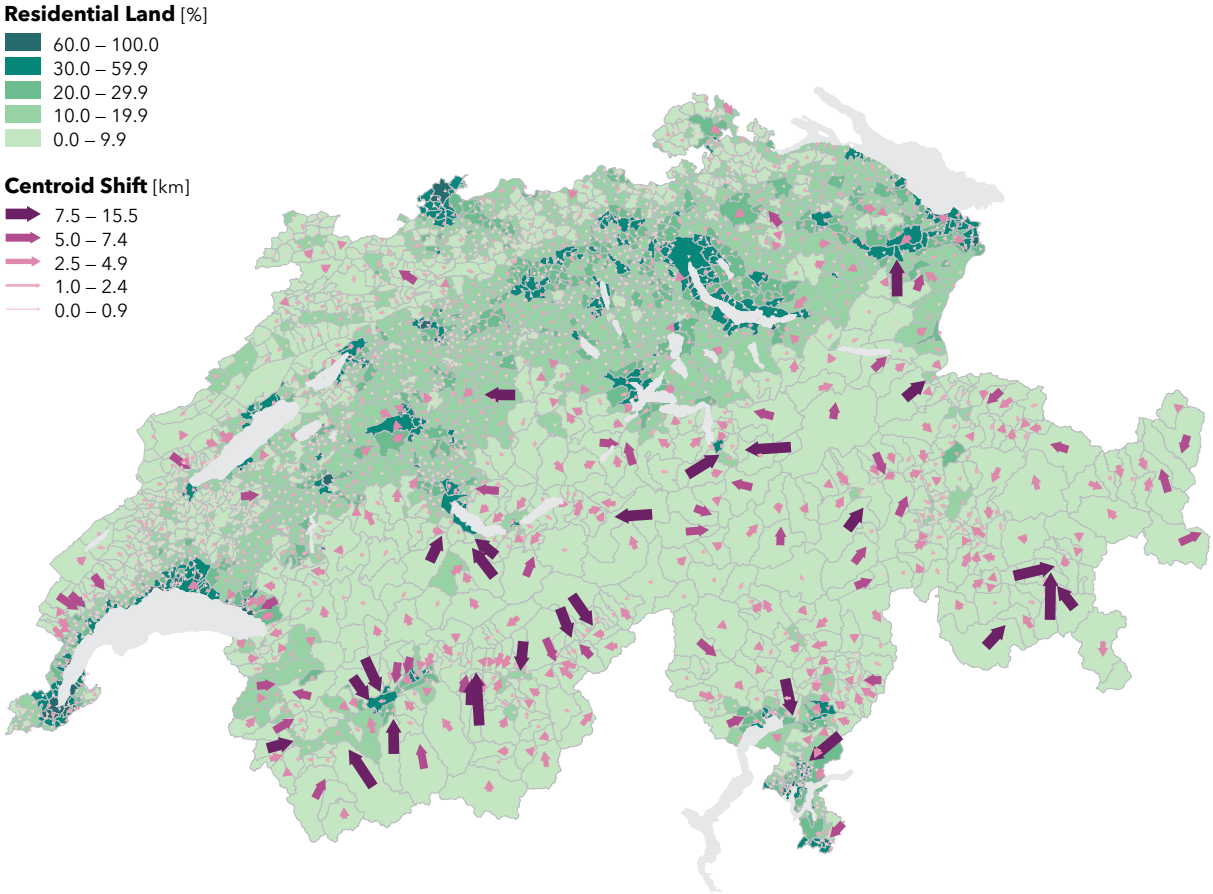

Figure 8. Effects of binary dasymetric refinement of enumeration units – changes of spatial extent (part of residential land) and centroid displacements (shift to the centroid of the residential land). The data is classified according to the fixed classes classification method. The figure is adapted from Boo et al. (2018b).

Figure 9 features two maps of human population density at the municipal level – the first one was computed based on the spatial extent of municipal units (**Figure 9A**) and the second one was based on the spatial extent of dasymetrically refined units (**Figure 9B**). For better visual comparison, we portrayed the recomputed population densities after dasymetric refinement using the same spatial units in a choropleth fashion.

The use of dasymetrically refined units yields substantially higher human population densities because the spatial extent for density recomputation is reduced to the portion of residential land within municipalities. This effect is remarkable in mountainous regions and in most municipalities of the flat

Central Plateau, where human population densities also perceptibly increase. *Human Population Density* recomputed with dasymetrically refined units was therefore expected to produce more accurate geographic distributions of this variable and more robust statistical associations in the models of canine cancer incidence rates.

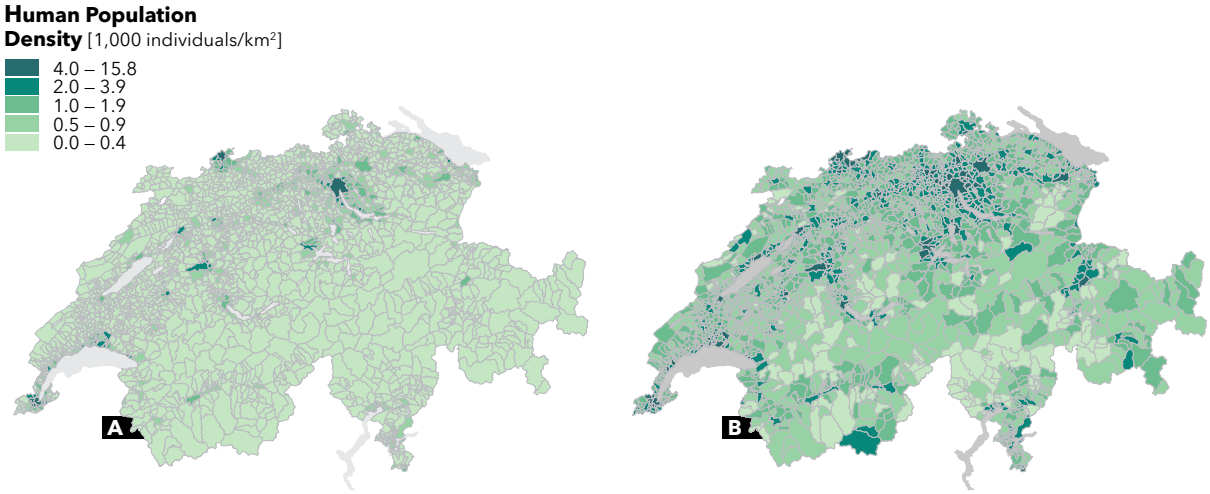

Figure 9. Human population density indicators resulting from the two enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. Both indicators are presented in a choropleth fashion. The data is classified according to the quantile classification method applied to the dasymetrically refined unfits. The figure is adapted from Boo et al. (2018b).

Figure 10 shows two maps of distance to the closest veterinary service averaged at the municipal level – the first was computed based on the spatial extent of the municipal units (**Figure 10A**), and the second on the spatial extent of dasymetrically refined units (**Figure 10B**). For better visual comparison, both spatial units are depicted in a choropleth fashion.

Despite the change of support, the averaged distances to veterinary care are surprisingly similar in both maps. This result may simply be explained by the fact that the Swiss road network is typically less developed in scarcely populated regions than in the densely populated Central Plateau. As a consequence, the differences resulting from the two enumeration types were negligible. As shown in **Table 4**, averaging the distance grid at the municipal level produced a median distance to veterinary care of 3.0 km (IQR=2.9), which is only slightly higher than the median distance of 2.7km (IQR= 2.8) after dasymetric refinement. In both cases, the relatively large IQRs suggest a persisting impact of large distances to veterinary services, typically in the Alps

(South) and in Jura Mountains (North-West). For this reason, we did not expect that *Distance to Veterinary Care,* recomputed using dasymetrically refined units, would lead to more accurate geographic distributions of this variable. As such, statistical associations in the models of canine cancer incidence rates were not expected to change or be more robust.

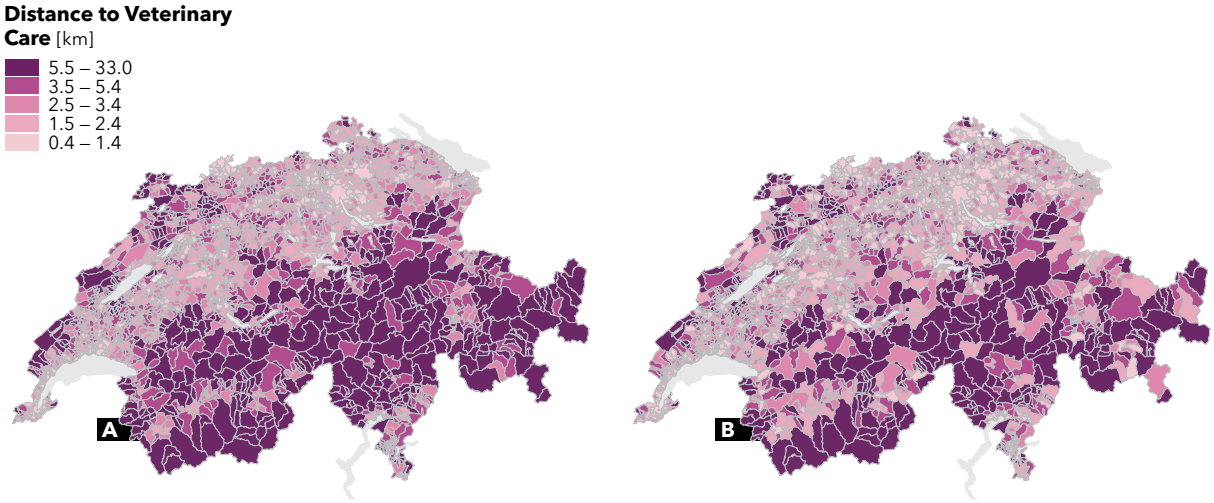

Figure 10. Distance to veterinary care indicators resulting from the two enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. Both indicators are presented in a choropleth fashion. The data is classified according to the quantile classification method applied to the dasymetrically refined unfits. The figure is adapted from Boo et al. (2018b).

Exploring uncertainty in the Swiss Canine Cancer Registry data

Figure 11 shows the geographic distribution of observed canine cancer incidence rates at the municipal level in Switzerland for the year 2008, as fit with the regression models. The rates seem to exhibit a particular geographic configuration, with higher rates in the German-speaking northeast of the Central Plateau (North-East) compared with the low-to-mixed rates in the French-speaking part of the country (South-West). In the Alps (South) and Jura Mountains (North-West), the rates were mostly very low or even zero, possibly, because of extreme fluctuations owing to the small sample sizes in these scarcely populated regions. **Figure 11** also provides visual support for our dasymetric framework – the incorporation of residential land facilitated a more accurate interpretation of the geographic distribution of the canine cancer incidence rates. For example, the physical proximity between

settlements, with potentially similar demographic and environmental characteristics, seemed to be a driver for similar rates.

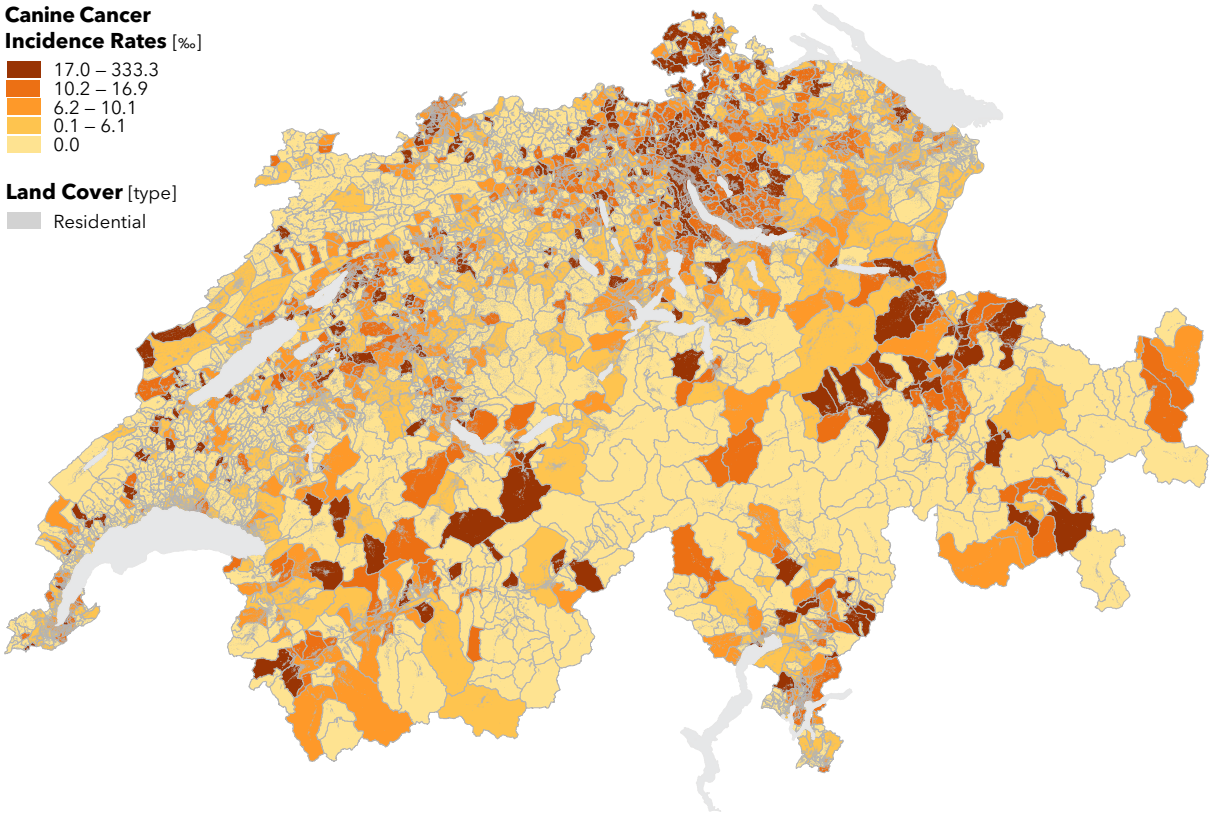

Figure 11. Geographic distribution of the canine cancer incidence rates in Switzerland in 2008. The data is classified according to the quantile classification method. Residential land is overlaid on the choropleth map. The figure is adapted from Boo et al. (2018b).

To garner insight into uncertainty within the SCCR data, we considered the four regression frameworks for count data – (1) Poisson model, (2) Poisson model with zero-inflation extension, (3) negative binomial model, and (4) negative binomial model with zero-inflation extension – for the two types of enumeration units (i.e., municipal units and dasymetrically refined units). The zero-inflated extensions implemented only *Average Age* because this is the only significant (P<.05) independent variable in the logistic model component. **Table 5** lists the AIC measures for the four regression models and both types of enumeration units.

Table 5. AIC measures for the different regression models based on the two enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. The table is adapted from Boo et al. (2018b).

Regression model	**(A) Municipal units**	**(B) Dasymetrically refined units**
(1) Poisson	6449.3	6419.7
(2) Negative binomial	5930.2	5910.7
(3) Poisson with zero inflation	6243.2	6223.5
(4) Negative binomial with zero-inflation	5894.5	5878.2

Table 6 presents the results of the pairwise relative-likelihood assessments and the likelihood-ratio tests to further determine the improvement of one model over another. The relative likelihood indicates the probability that the model with the lowest AIC minimizes the estimated information loss. In the likelihood ratio test, a positive χ^2 value rejects Model 1. The significance level of the test is reported in parentheses. The results of this assessment suggested that the negative binomial model with zero-inflation extension (**4**) outperformed the other models for both types of enumeration units. This is because these models had the lowest AIC measure.

Table 6. Pairwise relative-likelihood (RL) assessments and likelihood-ratio tests (χ^2) comparing the different regression models based on the two enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. The table is adapted from Boo et al. (2018b).

		(A) Municipal units		**(B) Dasymetrically refined units**	
Model 1	**Model 2**	**RL**	**χ^2 (*P*)**	**RL**	**χ^2 (*P*)**
(1) Poisson	**(2)** Negative binomial	.00	521.1 (*P*<.001)	.00	511.0 (*P*<.001)
(1) Poisson	**(3)** Poisson with zero inflation	.00	210.1 (*P*<.001)	.00	200.2 (*P*<.001)
(2) Negative binomial	**(3)** Poisson with zero inflation	.00	311.0 (*P*<.001)	.00	310.8 (*P*<.001)
(2) Negative binomial	**(4)** Negative binomial with zero-inflation	.00	39.7 (*P*<.001)	.00	36.5 (*P*<.001)
(3) Poisson with zero inflation	**(4)** Negative binomial with zero-inflation	.00	350.7 (*P*<.001)	.00	347.3 (*P*<.001)

The pairwise relative-likelihood assessments further confirmed this improvement as they reflected that the possibility that the other models could compete in minimizing the information loss of the negative binomial model with the zero-inflation extension being extremely low (i.e., .00). The pairwise likelihood-ratio tests also confirmed that this was a significant improvement with a resulting χ^2 of 350.7 (P<.001) and 347.3 (P<.001).

Table 7 presents the coefficient estimates, *P*-values, and the percentage of deviance reduction for each independent variable for the negative binomial model with the zero-inflation extension using both types of enumeration units.

The coefficient estimates suggested that *Average Age* involved negative relationships, specifically such that for each increasing year of age, the incidence rates decreased by 17.3% (municipal units) and 18.9% (dasymetrically refined units). Conversely, *Females per Male* and *Mixed-Breed Ratio* both yielded positive relationships, indicating that for each increasing percentage unit of females and mixed-breed dogs, the incidence rates rose by 1.0% (municipal units) and 2.0% (dasymetrically refined units) along with 3.1% (municipal units) and 2.0% (dasymetrically refined units), respectively.

Average Income Tax and *Human Population Density* both had positive relationships – for each 1,000 CHF per capita and 1,000 individuals per square kilometer, the incidence rates increased by 11.6% (both for municipal units and dasymetrically refined units) along with 4.1% (municipal units) and 8.3% (dasymetrically refined units), respectively. However, the latter association was not significant in the model based on municipal units (P=.23). *Distance to Veterinary Care* exhibited negative relationships, providing evidence that for each kilometer of distance to veterinary care, the incidence rates decreased by 3.0% (municipal units) and 3.9% (dasymetrically refined units).

Lastly, in the zero-inflation extension of the models, *Average Age* featured negative relationships, suggesting that for each increasing year, the odds of having zero incidence rates decreased by 97.3% (municipal units) and 97.5% (dasymetrically refined units).

Table 7. Coefficient estimates, P-values, and percentages of variance reduction (η^2) for the (**4**) negative binomial model with zero-inflation based on the enumeration types – (**A**) municipal units and (**B**) dasymetrically refined units. The table is adapted from Boo et al. (2018b).

	(A) Municipal units			(B) Dasymetrically refined units		
Coefficient	**Estimate**	***P***	**η^2 (%)**	**Estimate**	***P***	**η^2 (%)**
Negative Binomial						
Average Age	−0.19	<.001	20.40	−0.21	<.001	13.86
Females per Male	0.01	<.001	3.76	0.02	<.001	5.18
Mixed-Breed Ratio	0.03	<.001	20.40	0.02	<.001	18.39
Average Income Tax	0.11	<.001	21.53	0.11	<.001	23.06
Human Population Density	0.04	.23	0.75	0.08	<.001	9.33
Distance to Veterinary Care	–0.04	<.001	8.76	−0.03	<.001	4.92
Zero inflation (Logistic)						
Average Age	−3.61	<.001	24.41	−3.69	<.001	25.26

Table 7 also offers insights into how statistical performance is related to the effects of spatial data aggregation because it shows results for both municipal units and dasymetrically refined units. As mentioned, the coefficient estimate for *Human Population Density* is not significant (*P*=.23) in the model based on municipal units, and also involves a lower percentage of deviance reduction. On the contrary, the coefficient estimate for *Distance to Veterinary Care* is significant (*P*<.001) in the regression models based on both types of enumeration units, and the percentage of deviance reduction even rises when using municipal units. The table also indicates that the recomputation of *Human Population Density* and *Distance to Veterinary Care* influenced other independent variables. In particular, we observed higher percentages of deviance reduction for the independent variables accounting for confounding factors associated with potential underascertainment of cancer cases when using dasymetrically refined units.

Lastly, we compared the statistical performance of the regression models based on the two enumeration types through the relative-likelihood and the likelihood-ratio test. The former shows that the regression model based on municipal units is very unlikely (i.e., .00) to compete with the one employing dasymetrically refined units. The latter shows that the regression using dasymetrically refined units is a significant improvement (χ^2=16.3; *P*<.001).

3.2.4. Summary and key findings

This case study explores the effects of spatial data aggregation on models of canine cancer incidence rates. This was carried out through a systematic evaluation of changes in the statistical performance resulting from the dasymetric refinement of municipal units to their portion of residential land. To this end, two modeling scenarios based on different spatial units are contrasted – the first is based on municipal units, while the second on dasymetrically refined units. The two modeling scenarios involve the recompilation of density and distance variables implemented in the models.

- The negative binomial model with zero inflation extension outperforms the other models because it is designed for fitting the excess of structural zeros in the SCCR data;
- Dasymetric refinement produces more accurate geographic distributions and more robust statistical associations for the independent variable related to the density indicator (i.e., *Human Population Density*);
- Dasymetric refinement does not lead to more accurate geographic distributions or more robust statistical associations for the distance indicator (i.e., *Distance to Veterinary Care*); and
- Overall, using dasymetrically refined units to compute the independent variables has led to more robust statistical associations and enhanced statistical performance in the model of canine cancer incidence rates.

The results of this case study will be interpreted and discussed in Section 4.1.2. The next section presents the third case study of this thesis. It is concerned with the challenges and limitations associated with the effects of spatial non-stationarity and geographic scale.

3.3. INFLUENCES OF SPATIAL NON-STATIONARITY AND GEOGRAPHIC SCALE

This section addresses the third research question (RQ 3).

RQ 3 How does the selected *analytical framework* impact the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This is conducted through a case study published by Boo et al. (2018a), which reports original research performed by the author of this thesis. Gianluca Boo processed the data, developed and implemented the study design, interpreted the results, and wrote the first draft of the manuscript. Stefan Leyk and Christopher Brunsdon edited the manuscript, contributed to the design, implementation, and interpretation of the results. Sara I. Fabrikant and Andreas Pospischil edited the manuscript and contributed to the interpretation of the results. Ramona Graf collected and pre-processed the SCCR data. The content of the original manuscript is reported in a slightly altered form to better fit into the structure of this thesis.

3.3.1. Rationale

Inconsistent statistical associations may be linked to the effects of spatial non-stationarity (Fotheringham et al. 1996; Brunsdon et al. 1996) and geographic scale (Atkinson and Tate 2000; Tate and Atkinson 2001). To advance the understanding of these effects in models of canine cancer incidence rates, we designed an analytical framework inspired by the concept of regional models. This concept was recently proposed for robust analysis and diagnostic of spatial non-stationarity and aggregation effects in epidemiologic and demographic contexts (Leyk et al. 2012a; Maclaurin et al. 2015).

An essential characteristic of regional models is that they keep the structure of the conventional regression model unaltered because effects of spatial non-stationarity and geographic scale are implicitly embodied through the region to which the regression model is fit (Leyk et al. 2012a; Maclaurin et al. 2015). This results in a relatively simple modeling framework that, unlike

existing local models, does not incorporate uncertainties associated with the specification of spatial weights (Tiefelsdorf 2006; Cho et al. 2010).

We defined multiple regions based on a set of nearest-neighboring municipal units. Each region featured a specific central municipal unit and geographic scale, in other words, the number of nearest-neighboring municipal units. Regional models were then fit to regions involving all possible centers and geographic scales. Selected model diagnostics were next computed, summarized, and visualized through value-by-alpha maps (Roth et al. 2010) and scalograms (Dykes and Brunsdon 2007). The visual representations were examined to contrast the regional models with the conventional regression model.

Such a comparative assessment permitted us to uncover effects of spatial non-stationarity and geographic scale in the model of average canine cancer incidence rates. This finding provided insights into the choice of more appropriate modeling methods for capturing spatial structure in the spatial epidemiology of canine cancer as the two elements appeared to be highly connected.

3.3.2. Materials and methods

Canine cancer diagnostic examinations and demographic indicators

The SCCR consists of canine cancer diagnostic examinations collected retrospectively in Switzerland between 1955 and 2013. The diagnostic examinations involved necropsy, biopsy, and cytology tests at the veterinary hospitals of Zurich and Berne, as well as at a private laboratory located in the Zurich area (Grüntzig et al. 2015). The two veterinary hospitals own the only official laboratories for animal cancer diagnosis in Switzerland. The different diagnostic laboratories perform examinations based on cases submitted by veterinary services located across the entire country (Grüntzig et al. 2015, 2016).

Based on the residential addresses stored in the diagnostic data, we computed canine cancer incidence at the municipal level. This was done by linking the residential postcode to the unique identification number of the Swiss municipal units (SFSO 2017). For each municipal unit, the incidences

were computed on a yearly basis for the period 2008–2013 and then summed over the six years. Over this period, 20,209 new cancer cases were recorded in Switzerland with a median yearly value of 3,350 and an IQR value of 127. Despite the relative stability of the yearly incidence at the country level, they varied greatly at the municipal level, with 28% of the municipal units having a median value equal to or even lower than the IQR. Such an important local variability justifies the aggregation of the canine cancer incidence across six years to avoid spurious results associated with temporal variability. All types of malignant tumors were considered cancer cases, and dogs diagnosed with more than one cancer where considered single cases.

We also accessed the Swiss Canine Population Census, which is compiled by Animal Identity Service (ANIS) AG following the legal obligation for dog microchipping and registration established in Switzerland in 2006 (ANIS 2017). Based on the residential address of the registered dogs, we retrieved the number of at-risk dogs at the municipal level on a yearly basis for the period 2008–2013. No exclusion criterion as to age or sex was adopted. We also aggregated the population counts for each municipality over the six years to avoid extreme fluctuations based on sample variability (Elliott et al. 1996; Beale et al. 2008).

As a result of the total number of canine cancer cases and the population counts recorded within municipalities over the six years, we were able to compute the average canine cancer incidence rates for the period 2008–2013. We derived additional indicators associated with known demographic risk factors for several canine cancers (Merlo et al. 2008; Brønden et al. 2010; Dobson 2013). These indicators are average age (in months), females per male (in percent), and average weight (in kilograms) of dogs within the different municipal units each year, combined during the period 2008–2013.

Indicators of potential underascertainment of canine cancer cases

We computed three additional indicators for potential underascertainment of cancer cases, a confounding factor known to affect the study of canine cancer registry data (Brønden et al. 2007; Nødtvedt et al. 2011).

The first indicator estimated the urban character of municipalities, as lower levels of underascertainment of cancer cases are expected to occur in urban locations where veterinary check-ups are typically more frequent (Gavazza et al. 2001; Bartlett et al. 2010). For this purpose, we computed dogs per capita (in percent) across municipalities using the Swiss Canine Population Census data (ANIS 2017) and the Swiss Federal Statistical Office census data (SFSO 2017) for the period 2008–2013. The reason is that different characteristics such as the status of the dog (i.e., companion versus working) and the type of households (i.e., smaller versus larger) influence the number of dogs per capita living in urban and rural municipalities (Pospischil et al. 2013).

Secondly, we also considered that wealthier municipalities had reduced levels of underascertainment of cancer cases as well, because of the availability of financial means for regular veterinary check-ups (Bartlett et al. 2010; O'Neill et al. 2014). Therefore, we calculated average federal income tax (in 1,000 CHF per capita) by normalizing income tax information collected by the Swiss Federal Tax Administration (SFTA 2017) and the Swiss Federal Statistical Office census data (SFSO 2017) for the period 2008–2012. We could not access federal income tax information for 2013 because the data was not publicly available at the time of this study.

Lastly, we further addressed the frequency of regular veterinary check-ups by computing distance to veterinary care (in kilometers) within municipal units. This was performed by creating a hectometric raster (i.e., with a 100m x100m resolution) representing the distance to veterinary services along roads and averaging the raster values within those municipal units, based on the location of their centroid (Delamater et al. 2012; Bliss et al. 2012). The raster was created using the addresses of the 938 veterinary services registered in the official Swiss Yellow Pages online database in 2014 (Swisscom Ltd. 2017). The Swiss road network for 2014 was obtained as vector data from the VECTOR25 data model of the Swiss Federal Office of Topography (SFOT 2017).

We could not access data pertaining to the addresses of veterinary services for previous years because such historical information was not readily available to us. However, information issued by the Swiss Registry of Medical

Professions confirms that changes in the number of licensed veterinarians within this period are negligible (FOPH 2017).

Table 8. Statistical distribution of the independent variables employed in the conventional regression model. The table is adapted from Boo et al. (2018a).

Variable	Median	IQR	Min	Max
Average Age (months)	81.9	13.7	47.7	138.0
Females per Male (percent)	51.3	6.6	0.0	83.7
Average Weight (kilograms)	22.6	3.7	8.2	41.3
Dogs per Capita (percent)	13.2	8.0	1.8	276.0
Average Income Tax (1,000 CHF per capita)	0.6	0.5	0.1	30.3
Distance to Veterinary Care (kilometers)	3.0	2.9	0.4	33.0

Modeling average canine cancer incidence rates

We fit the average canine cancer incidence rates using a Poisson regression framework as this is one of the most typical methods for modeling incidence and rates (Frome 1983; Frome and Checkoway 1985).

In so doing, we relied on the assumption that the data was Poisson distributed, in particular having the property whereby the conditional variance is equal to the conditional mean (Cameron and Windmeijer 1997). However, mild violations of this assumption have often been described and accepted (Cameron and Trivedi 1990). Given the purpose of our case study, we do report the results of the over-dispersion test (α=.05), but we did not consider alternatives to the Poisson model (Cameron and Trivedi 1986). This was because our goal was to examine model diagnostics rather than test different modeling frameworks (Berk and MacDonald 2008).

As the Poisson model is essentially designed for count data, we first fit the observed canine cancer cases between 2008 and 2013 (y) through the following independent variables (x): *Average Age* (in months), *Females per Male* (in percent), *Average Weight* (in kilograms), *Dogs per Capita* (in percent), Average *Income Tax* (in 1,000 CHF per capita), and *Distance to Veterinary Care* (in kilometers), according to **Equation 5**. The fit canine cancer incidence ($\hat{y}$) were then adjusted according to the at-risk canine population (in number of individuals) between 2008 and 2013 (e) and log-transformed, hence computing average canine cancer incidence rates for the period. In the

equation, α is the intercept, β the multiplicative coefficient estimated for each independent variable, and ε the error term (Frome 1983; Frome and Checkoway 1985).

$$\log(\hat{y}(y|x)) = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \log(e) + \varepsilon \quad \textbf{(Equation 5)}$$

To assess the statistical performance in our baseline model, we examined various diagnostics about the direction, multiplicative effect (i.e., exp(ß), in percent), and significance level (α=.05) associated with the estimated model coefficients (Frome 1983; Frome and Checkoway 1985). When computing the significance levels, we considered robust standard errors to account for mild deviations from the Poisson distribution (Cameron and Trivedi 1986). We also tested the independent variables for multicollinearity to detect correlations among the independent variables as this may introduce problems in the estimation of the model coefficients (Gujarati and Porter 2003). For this purpose, we employed the variance inflation factor (VIF) and reported its square root value (SQRVIF). A SQRVIF greater than 2.0 indicates critical multicollinearity (Fox 2015).

We then evaluated whether our baseline model provided a significant (α=.05) improvement over the null model, that is, the model with the intercept only. As such, we performed a likelihood ratio test (Lewis et al. 2011) and reported the Chi-squared statistic (χ^2) (Neyman and Pearson 1933). To assess the statistical performance of the models, we computed the $R^2_{McFadden}$ statistic (McFadden 1973). Similar to the likelihood ratio test, the $R^2_{McFadden}$ statistic evaluates the improvement of the baseline model over the null model considering the explained variability. As with the standard R-squared statistic, as a $R^2_{McFadden}$ statistic approaches 0, it indicates a lower model fit; a value of 1 indicates a perfect model fit (Cameron and Windmeijer 1996). In practice, the $R^2_{McFadden}$ statistic is more conservative, and the respective values are lower than standard R-squared values. Values between 0.2 and 0.4 suggest an excellent model fit (Domencich and McFadden 1975).

Investigating spatial non-stationarity and geographic scale

To build the regional models, we fit the baseline model presented earlier within multiple regions based on a set of nearest-neighboring municipal units (Leyk et al. 2012a; Maclaurin et al. 2015). Firstly, we defined the modeling regions by considering every municipal unit as a center. Secondly, based on the Euclidean distance between the different centers, we iteratively selected nearest neighboring units spanning from one to the total number of municipal units within the study area (Lloyd 2010). These steps allowed us to define the multiple regions as a function of their centers and the number of nearest neighboring municipal units.

On the one hand, this enabled us to fit models to each of the regions, thus assessing potential spatial non-stationarity in estimated relationships across regions. On the other hand, we were also able to establish the effects of geographic scale – estimated by the number of nearest neighboring municipal units involved in the regions – on these statistical associations. However, as the geographic scale decreases, sample-size effects become critical to the regional models. For this reason, we enforced a minimum number of nearest neighboring municipal units to ensure acceptable statistical power (ß=.80) given a standard significance level (α=.05) and a small effect size (f^2=.04) (Ferguson 2009).

We contrasted the regional models with the diagnostic tools presented before by assessing potential changes in the direction and multiplicative effect of significant model coefficients (α=.05) (Frome 1983; Frome and Checkoway 1985) as well as in the relative statistical performance of the model (Neyman and Pearson 1933; McFadden 1973). To facilitate this comparative task, we computed summary statistics for the diagnostics of the different regional models. The summary statistics were classified into quartiles to produce robust measures of central tendency (i.e., the median) and spread (i.e., IQR) across the multiple diagnostics (Wan et al. 2014).

At this point, we mapped the geographic distribution of both median and IQR measures for the regional models, based on the location of the regional centers. In so doing, we built value-by-alpha maps to simultaneously depict median values through a standard continuous color

scale and IQR values through variations in the alpha parameter, or the opacity level (Roth et al. 2010). This technique was meant to grant a first insight into potential impacts of spatial non-stationarity and geographic scale across the multiple regional models. To further investigate the effects of geographic scale, we also examined scalograms, a graphic technique to assess changes in the model diagnostics across the different nearest neighboring municipal units employed to define the regions (Dykes and Brunsdon 2007). On the y-axis of the filled-area plots, we present the summary statistics according to the quartile classification method, and on the x-axis, we indicate the number of nearest neighboring municipal units characterizing the regional models.

3.3.3. Results

Fitting the global regression model

Figure 12 portrays the geographic distribution of the observed average canine cancer incidence rates for the period 2008–2013 across all Swiss municipal units as fit with the conventional regression model. Overall, the average canine cancer incidence rates manifest strong regional patterns, dominated by higher rates in the municipal units located in the eastern part of the country, across the Cantons of Zurich and Schaffhouse (North-East), the Canton of Grisons (East) and the Canton Ticino (South-East). We identified additional regional patterns associated with a rural-urban cleavage. Municipal units belonging to the major urban agglomerations exhibited substantially greater rates than the rural hinterland, namely, the Cantons of Vaud, Fribourg and Berne (West), the Alps (South), and the Jura Mountains (North-West).

Fitting the baseline model globally, through a conventional regression model, resulted in a likelihood-ratio test statistic of χ^2=3,878.6 (P<.001), confirming a significant improvement over the model with the intercept only. The $R^2_{McFadden}$ statistic was 0.20, suggesting a relatively robust statistical performance. The overdispersion test returned a value of 4.3 (P<.001), indicating significant overdispersion.

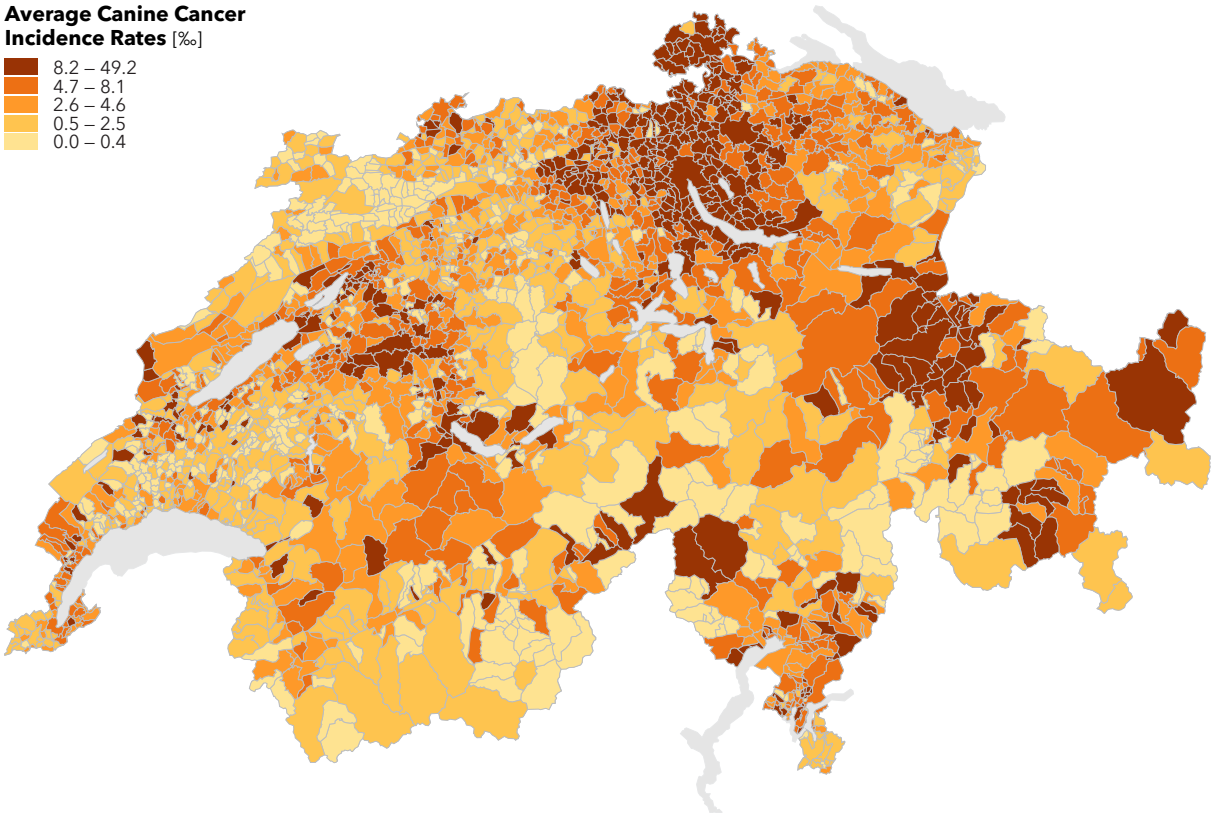

Figure 12. Geographic distribution of the average canine cancer incidence rates in Switzerland for the period 2008–2013. The data is classified according to the quantile classification method. The figure is adapted from Boo et al. (2018a).

Table 9 shows that all model coefficients are statistically significant ($P<.05$), and the SQRVIF values are consistently below 2.0, indicating the absence of critical multicollinearity. Demographic risk factors, such as *Average Age*, had a negative relationship, as for each increasing month of age, the average cancer incidence rates decreased by 1.9%. Conversely, both *Females per Male* and *Average Weight* exhibited positive relationships – for each increasing percentage unit of females and each increasing kilogram, the average cancer incidence rates rose by 3.0% and 3.9%, respectively. Confounding variables accounting for potential underascertainment of cancer cases, such as *Dogs per Capita* and *Distance to Veterinary Care*, exhibited negative relationships – for each increasing percentage unit of dogs and kilometer of distance, the average cancer incidence rates decreased by 6.0% and 4.6%, respectively. Lastly, Average *Income Tax* exhibited a positive relationship – for each increasing 1,000 CHF per capita, the average cancer incidence rates rose by 9.4%.

Table 9. Coefficient estimates, lower and upper 95% CIs, P-values and SQRVIFs for the conventional regression model. The table is adapted from Boo et al. (2018a).

Coefficient	Estimate	Lower CI	Upper CI	P	SQRVIF
Average Age	–0.020	–0.024	-0.016	< .001	1.09
Females per Male	0.029	0.021	0.037	< .001	1.03
Average Weight	0.039	0.019	0.059	< .001	1.22
Dogs per Capita	–0.062	–0.075	–0.049	< .001	1.25
Average Income Tax	0.090	0.059	0.121	< .001	1.03
Distance to Veterinary Care	–0.047	–0.063	–0.031	< .001	1.12

Fitting the regional regression models

The power analysis of the conventional regression model returned a minimum sample size of 347 municipal units. As a consequence, after excluding the center, the set of nearest-neighboring municipal units defining the multiple regions could range between 346 and 2,324. Iterating through all possible regions produced 4,594,548 regional models. In each of these models, the likelihood-ratio test statistics indicated a significant (P<.05) improvement over the model with the intercept only. The overdispersion tests returned values between 2.0 and 6.3 (P<.05), suggestive of significant overdispersion. None of the regional models produced model coefficients featuring critical multicollinearity (SQRVIF<2.0), but occasionally non-significant (P>0.05) model coefficients were present. These were discarded when producing the summary statistics and visualizations.

Figure 13A shows the geographic variations in the $R^2_{McFadden}$ statistics through a value-by-alpha map. This indicated a clear trend in the median $R^2_{McFadden}$ measures, characterized by higher values in the center of the country, transitioning into lower values toward the East and the West. In the Western part of the country, we found very high IQRs, indicating a larger spread of $R^2_{McFadden}$ measures across geographic scales. Conversely, IQRs were closely centered around the median in the Central and Eastern parts of the country. **Figure 13B** depicts the variations in the $R^2_{McFadden}$ measures across geographic scales using a scalogram. On the one hand, for smaller numbers of nearest neighboring units, the $R^2_{McFadden}$ statistics exhibited a higher spread, spanning from low to high values. On the other hand, for larger numbers of

nearest neighboring units, the $R^2_{McFadden}$ measures had a reduced spread, becoming increasingly similar to the $R^2_{McFadden}$ statistic of the global model.

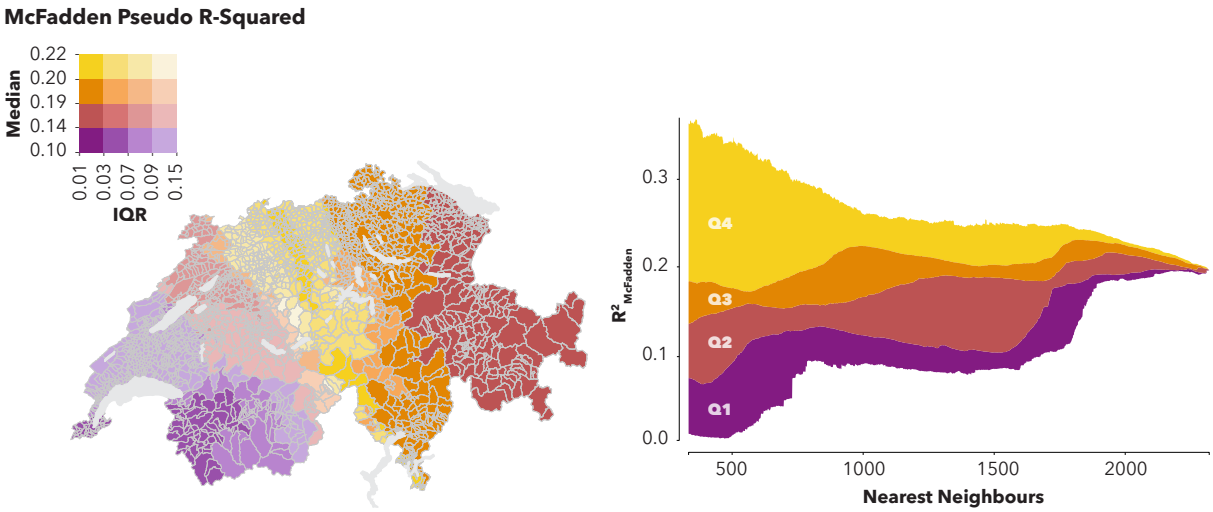

Figure 13. Variations of the $R^2_{McFadden}$ measures across (**A**) the center and (**B**) the geographic scale of the regional models. The data is classified according to the quantile classification method. The figure is adapted from Boo et al. (2018a).

Figure 14 shows the geographic variations in the multiplicative effects associated with significant coefficient estimates through value-by-alpha maps. These reveal clear trends in the median multiplicative effects, mostly across the East-West axis.

In the Western part of the country, *Average Age* (**Figure 14A**) and particularly *Average Weight* (**Figure 14C**), which even had contrasting median relationships, both presented negative median multiplicative effects in the Eastern part. *Females per Male* (**Figure 14B**) showed positive median multiplicative effects across the entire country. *Dogs per Capita* (**Figure 14D**) and *Distance to Veterinary Care* (**Figure 14F**) both showed negative median multiplicative effects while *Average Income Tax* (**Figure 14E**) exhibited positive median multiplicative effects. All multiplicative effects resulting from the significant coefficient estimates demonstrated relatively high levels of spread across geographic scales, with the highest IQRs reported for *Average Weight*, Average *Income Tax*, and *Distance to Veterinary Care*. Nonetheless, the effects of geographic scale did not seem to follow any specific geographic configuration.

Figure 15 portrays variations in the multiplicative effects associated with the significant coefficient estimates across geographic scales through

scalograms. These illustrate high spread in the multiplicative effects at smaller geographic scales, which transition into lower spread at increasing geographic scales.

Average Age (**Figure 15A**), *Females per Male* (**Figure 15B**), and *Average Weight* (**Figure 15C**) had the highest variability in terms of multiplicative effects, which also resulted in contrasting relationships. This suggested that independent variables accounting for demographic risk factors have both positive and negative multiplicative effects, depending on the geographic scale under consideration. Conversely, the independent variables accounting for confounding factor associated with potential underascertainment of cancer cases, such as *Dogs per Capita* (**Figure 15D**), *Average Income Tax* (**Figure 15E**), and *Distance to Veterinary Care* (**Figure 15F**), had more consistent relationships concerning geographic scale. Only sporadically did these independent variables exhibit both positive and negative multiplicative effects, evincing the important effects of geographic scale.

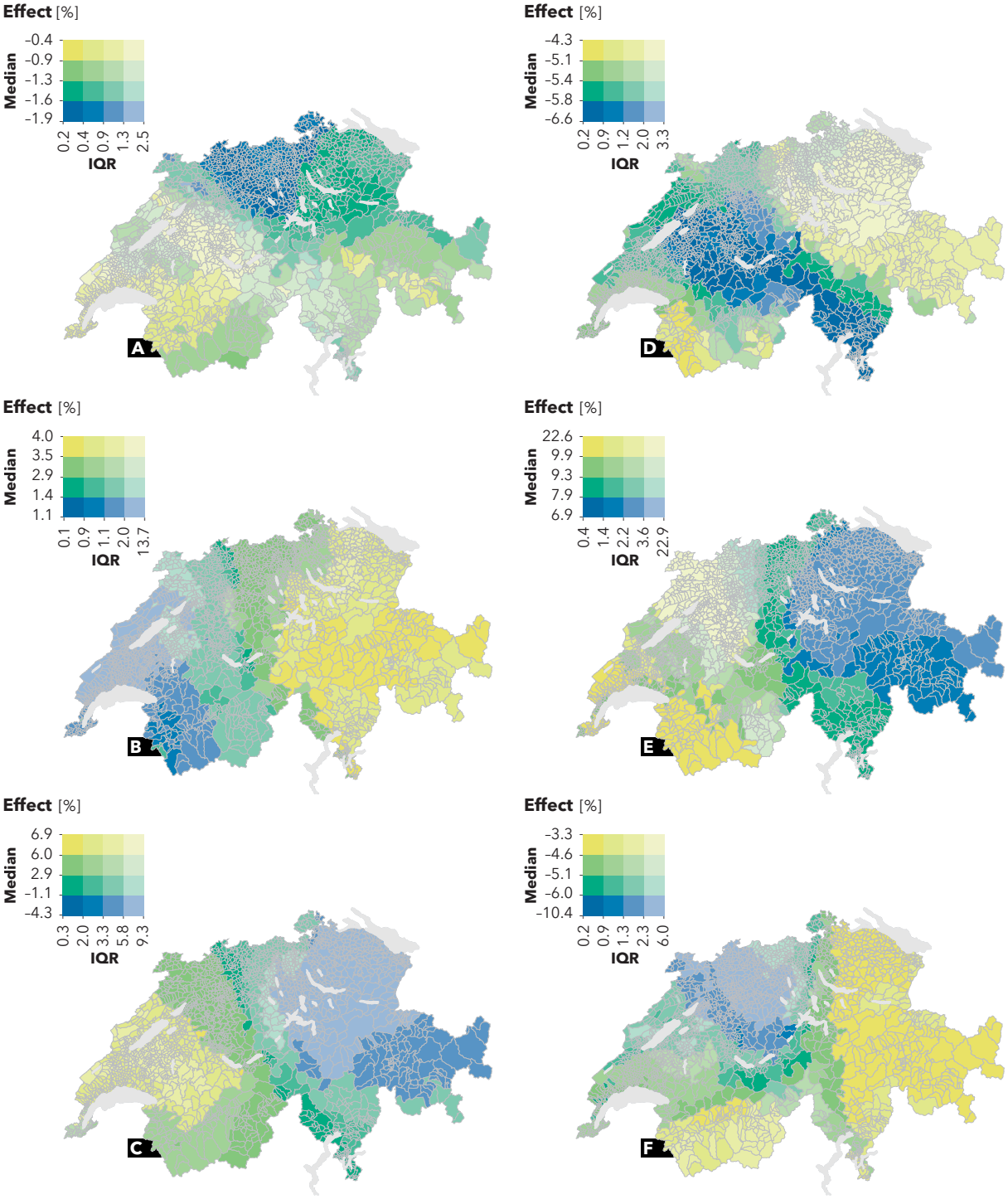

Figure 14. Variations of the multiplicative effects across the center of the regional models for (**A**) *Average Age* (in months), (**B**) *Females per Male* (in percent), (**C**) *Average Weight* (in kilograms), (**D**) *Dogs per Capita* (in percent), (**E**) *Average Income Tax* (in 1,000 CHF per capita), and (**F**) *Distance to Veterinary Care* (in kilometers). The data is classified according to the quantile classification method. The figure is adapted from Boo et al. (2018a).

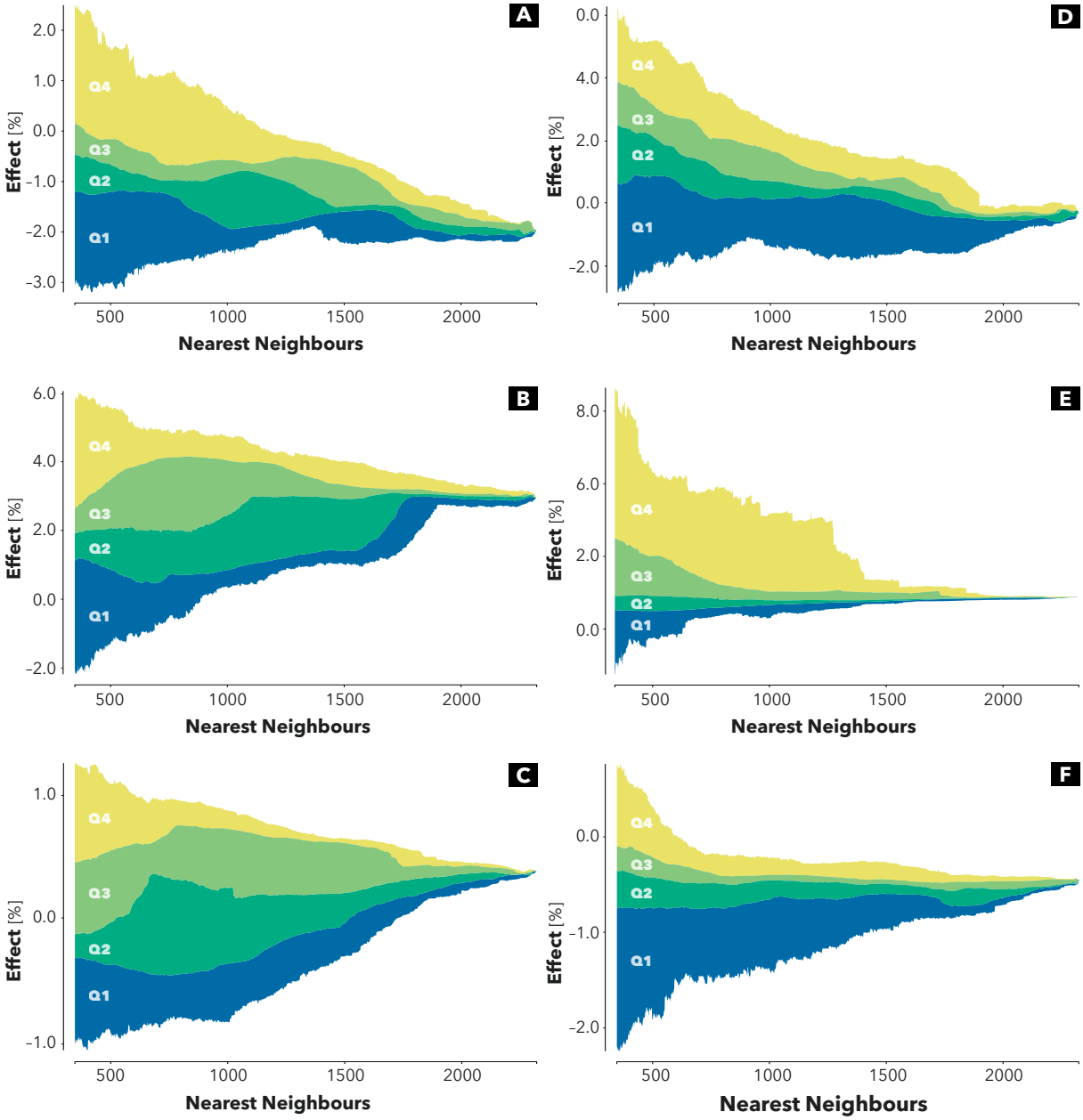

Figure 15. Variations of the multiplicative effects across the geographic scale of the regional models for (**A**) *Average Age* (in months), (**B**) *Females per Male* (in percent), (**C**) *Average Weight* (in kilograms), (**D**) *Dogs per Capita* (in percent), (**E**) *Average Income Tax* (in 1,000 CHF per capita), and (**F**) *Distance to Veterinary Care* (in kilometers). The data is classified according to the quantile classification method. The figure is adapted from Boo et al. (2018a).

3.3.4. Summary and key findings

This case study assesses the effects of spatial non-stationarity and geographic scale on a model of average canine cancer incidence rates. To determine these crucial methodical considerations, we fit average canine cancer incidence rates for the period 2008-2013 across Swiss municipal units through multiple regions. The regions were identified by their central municipal unit and geographic scale, in other words, the number of nearest-neighboring municipal units. Regional models were fit to regions involving all possible centers and geographic scales. Diagnostic summaries across the different regional models were then computed and contrasted with the diagnostics of the conventional regression model with value-by-alpha maps and scalograms.

- Canine cancer incidence rates were averaged over several years. However, municipal units with zero incidence persisted, typically, in the Alps and Jura Mountains;
- The statistical performance ($R^2_{McFadden}$) of the regional models varied considerably across the study area with a lower goodness-of-fit in the regions with lower average canine cancer incidence rates;
- The same independent variable could be associated with contrasting multiplicative effects when estimated within different regions, particularly for the variables related to demographic risk factors; and
- All independent variables were affected by geographic scale, to some extent. However, these effects did not seem to follow any specific geographic configuration.

The results of this case study will be interpreted and discussed in Section 4.1.3. The next chapter, Chapter 4, interprets the results of the three case studies with the objective of answering the proposed research questions.

4. DISCUSSION

This chapter discusses the three case studies reported in Chapter 3. The first section presents a general discussion interpreting the results of these studies, referring to the identified knowledge gaps. The second section considers the general limitations of the case studies from both an epidemiological and statistical modeling perspective. The third section revisits the previous discussions by providing an answer to the proposed research questions. Finally, the fourth section examines the connections between this thesis and the body of knowledge of GIScience.

4.1. GENERAL DISCUSSION

4.1.1. Underascertainment of cancer cases

The first case study aimed to address the first knowledge gap (KG 1).

KG 1 *Society and context* play an essential role in the spatial epidemiology of canine cancer (Dorn et al. 1968b). The *implications of underascertainment of cancer cases* and the effects on the statistical performance of models of canine cancer incidence need to be better understood for effective environmental-sentinel applications.

To fill this particular knowledge gap, the case study systematically compared two scenarios for modeling canine cancer incidence in a geographic-correlation study framework. The first scenario, depicted in **Figure 3A**, consisted of the complete enumeration of canine cancer incidence within all Swiss municipal units in 2008. The second scenario, depicted in **Figure 3B**, consisted of a filtered sample that systematically discarded municipal units presenting structural zeros. As a reminder, these are the municipal units where no diagnostic examination was performed because of the underascertainment of cancer cases (Boo et al.).

Such a filtering step demonstrated that structural zeros influence the statistical distributions of the cancer incidence data, as indicated by the robust measures of central tendency and spread (i.e., the median and IQR). These measures also suggested that the structural zeros produce an increased deviation from the assumed Poisson distribution (Cameron and Trivedi 1986, 1990). However, in spite of increased overdispersion in the data, no substantial deterioration in the statistical performance of the model of canine cancer incidence could be detected through the $R^2_{McFadden}$ statistics. Further, the goodness-of-fit did not appear to be affected by the presence of structural zeros.

Table 2 shows there to be no relevant change in the multiplicative effects associated with the different coefficient estimates. In this regard, *Average Age* had a persistent negative statistical relationship to canine cancer incidence, contradicting prior research on canine cancer (Bronson 1982; Eichelberg and Seine 1996). Still, this result could link to similar findings by Bonnett and Egenvall (2010), who suggested that such an association can be based, at least partially, on owner decisions about care. These decisions typically involve selective underascertainment of cancer cases, especially in very old dogs.

The theme of underascertainment of cancer cases also emerges in the significant statistical associations for the confounding factors, *Average Income*, *Human Population Density*, and *Distance to Veterinary Care*. **Table 2** also details how these independent variables had a higher proportion of variance reduction in the model of canine cancer incidence featuring structural zeros. This result indicates that the model might have captured the structural zeros through the suggested confounding factors, or stated differently, as an extreme manifestation of the underascertainment of cancer cases (Hu et al. 2011; He et al. 2014). The use of these independent variables is therefore crucial to the modeling effort.

The geographic distribution of the Pearson residuals depicted in **Figure 4** illustrates a general decrease in the predictive power of the model of canine cancer incidence associated with the presence of structural zeros (Cameron and Windmeijer 1996). Nevertheless, the municipal units featuring

structural zeros resulted in acceptable model over-estimations, with absolute residuals values below 2.0. For this reason, critical model over- and under-estimations were, at first, expected to be difficult to detect through the measures of average error (i.e., MAE and RMSE) produced using model cross-validation (Chai and Draxler 2014).

Nevertheless, as portrayed in **Figure 5**, model cross-validation enabled detection of an increased spread of coefficient estimates for nearly all independent variables in the model of canine cancer incidence featuring structural zeros. Moreover, *Females per Male* indicated contrasting statistical associations across model iterations, exhibiting both positive and negative multiplicative effects. The negative multiplicative effects contradict prior research on canine cancer suggesting that females have a higher cancer incidence (Bronson 1982; Michell 1999). In additions, less robust statistical associations across model iterations proved that structural zeros affected the statistical performance of the model of canine cancer incidence (Snee 1977; Picard and Cook 1984).

Table 3 shows that the model of canine cancer incidence featuring structural zeros also produced inflated measures of average error (i.e., MAE and RMSE), and the decrease in predictive power was only related to less than 5% of the errors. Cross-validation, therefore, facilitated discerning that structural zeros also impact the predictive power of the model of canine cancer incidence (Willmott 1981; Hyndman and Koehler 2006). In spite of these results, significant Spearman's ρ correlations between the RMSE and independent variables employed in the training sets could not relate to the sole presence of structural zeros.

Nevertheless, independent from the presence of structural zeros, a significant Spearman's ρ correlation was found between the average population sizes in the training set and RMSE, both for the complete enumeration and filtered sample, just as depicted in **Figure 6**. This was an expected result highlighting the importance of sample size in statistical modeling because larger samples typically feature higher statistical power (Cohen 1992). Higher statistical power involves, in turn, a greater likelihood

of successfully determining an effect, hence the more robust statistical associations when filtering of structural zeros (Lawson 2006).

In summary, the cross-validation framework enabled discovering substantial effects of structural zeros on the statistical performance and predictive power of the model of canine cancer incidence (Snee 1977; Picard and Cook 1984). Such an assessment also confirmed that sample size might be particularly critical to the generalization of the model of canine cancer incidence because of statistical power. For this reason, it is possible to contend that underascertainment of cancer cases introduces substantial issues of generalizability of the model of canine cancer incidence. However, the findings also demonstrate that generalizability can be addressed by developing a finer understanding of society and context because this knowledge can be successfully implemented within a modeling framework. This strategy is expected to be crucial to potential environmental-sentinel applications in Switzerland.

4.1.2. Spatial data aggregation

The second case study aimed at addressing the second knowledge gap (KG 2).

KG 2 The spatial epidemiology of canine cancer is usually conducted at the individual level (O'Brien et al. 2000). For potential environmental-sentinel applications, it is necessary to better understand the *effects of spatial data aggregation* of canine cancer cases and potential explanatory variables on the statistical performance of models of canine cancer incidence.

To fill this particular knowledge gap, the second case study further assessed canine cancer incidence within a geographic-correlation study framework by focusing on the issue of spatial data aggregation. This was carried out by investigating potential improvements in statistical performance associated with the use of independent variables computed based on dasymetrically refined enumeration units in contrast with standard municipal units (Boo et al. 2018b). Such an enumeration-unit refinement involved the concept of binary dasymetric mapping, a cartographic technique established

by Wright (1936) to produce more accurate distributions of areal data considering the geographic context (Eicher and Brewer 2001; Mennis 2009).

For this study purpose, two kinds of independent variables — *Human Population Density* and *Distance to Veterinary Care* — were examined to inform of the potential effects of spatial data aggregation (Cressie 1996; Cressie et al. 2009). This assessment was meant to underscore general issues in computing density and distance indicators within the specific Swiss context. The reason was that variations in local geographic characteristics determine where habitable and uninhabitable land can be found across the country. Therefore, the areal extent and geographic distribution of the residential land were expected to vary considerably between mountainous and flat regions (Vega Orozco et al. 2015). Given these particular geographic settings, summarized in **Figure 8**, a simple binary dasymetric refinement was expected to produce more accurate summary statistics, especially in mountainous regions.

As seen in **Figure 9**, computing the density indicator based on the spatial extent of dasymetrically refined units produced more accurate geographic distributions of *Human Population Density* when compared to the same indicator computed with the spatial extent of municipal units. Nevertheless, when computing the distance indicator, the geographic distributions of *Distance to Veterinary Care* depicted in **Figure 9** would remain similar across the two enumeration units despite the change in spatial support resulting in different spatial extents.

These preliminary results confirmed, on the one hand, known benefits of dasymetric refinement for computing population density distributions (Eicher and Brewer 2001; Mennis 2009). On the other hand, dasymetric refinement did not produce more accurate distributions of the distance indicators considering potential change of support problems (Cressie 1996; Cressie et al. 2009). The reason for that is most likely the specific characteristics of the Swiss road network, which is often far less developed in mountainous regions than in the flat and densely populated Central Plateau (Erath et al. 2008).

When contrasting the different regression models of canine cancer incidence rates through the pairwise assessments presented in **Table 5** and

Table 6, the negative binomial models with zero-inflation extension consistently exhibited the best statistical performance. This result further confirmed the excess of structural zeros in canine cancer incidence retrieved from the SCCR in 2008. The reason for that is the zero-inflation component of the model is explicitly designed to accommodate an excess of structural zeros (Zeileis et al. 2008; Hu et al. 2011).

The coefficient estimates of the negative binomial models with the zero-zero inflation extension presented in **Table 7** shows that the negative binomial component produced similar multiplicative effects across the two enumeration types. These multiplicative effects were in agreement with those of the first case study (Boo et al.). Also, *Mixed-Breed Ratio* had a positive statistical relationship with canine cancer incidence rates, contradicting prior research on canine cancer (Michell 1999; Proschowsky et al. 2003). This finding was still difficult to compare because mixed-breed dogs could have dramatically different life spans. Additionally, the zero-inflation extension of the models showed that *Average Age* had a negative multiplicative effect on the presence of structural zeros, thereby confirming the prior findings on increasing canine cancer incidence in older dogs (Bronson 1982; Eichelberg and Seine 1996).

Table 7 also provides insights into the potential effects of spatial data aggregation on the model of canine cancer incidence rates. Despite similar multiplicative effects, *Human Population Density* produced significant coefficient estimates only in the regression model employing dasymetrically refined units. This type of enumeration unit also featured a higher percentage of deviance reduction. These results confirmed that, besides computing more accurate geographic distributions, dasymetric refinement also produced more robust statistical associations for this density variable (Eicher and Brewer 2001; Mennis and Hultgren 2006).

In contrast, the coefficient estimate for *Distance to Veterinary Care* did not show any relevant change associated with dasymetric refinement. Furthermore, the percentage of deviance reduction even increased when using municipal units. This result indicated that, because of the specific characteristics of the Swiss road network, dasymetric refinement would not lead to more to more robust statistical associations of this distance indicator.

Such a result underscored the complexity of challenges connected with both with change of support problems and spatial data aggregation (Cressie 1996; Cressie et al. 2009).

Nevertheless, when comparing the statistical performance of the regression models based on the two enumeration types, the tests established an improvement associated with the use of dasymetrically refined units. In summary, these results suggest that dasymetric refinement should be considered to address the effects of spatial data aggregation in models of canine cancer incidence. Also, given that the ancillary data employed for dasymetric refinements (i.e., residential land) concerns specifically humans, the same refinement strategy could apply to potential environmental-sentinel applications in Switzerland.

4.1.3. Spatial non-stationarity and geographic scale

The third case study sought to deal with the third knowledge gap (KG 3).

KG 3 The spatial epidemiology of canine cancer usually considers small geographic scales (Tedardi et al. 2015). It is vital to underscore the *influences of spatial non-stationarity and geographic scale* on the statistical associations estimated in models of canine cancer incidence for potential environmental-sentinel applications.

To fill this particular knowledge gap, consistently within a geographic-correlation study framework, the third case study contrasted a conventional regression model with multiple regional models. The reason was to uncover the effects of spatial non-stationarity and geographic scale in models of average canine cancer incidence rates in Switzerland for the period 2008-2013 (Boo et al. 2018a). These modeling issues are difficult to detect in a conventional regression model and can only be assessed through local or regional models as they allow for geographic variations in model coefficients (Fotheringham and Brunsdon 1999; Lloyd 2010).

As witnessed with the $R^2_{McFadden}$ test, the conventional regression model produced a relatively important goodness-of-fit (Domencich and McFadden 1975). This statistical performance was in spite of significant overdispersion.

Furthermore, **Table 9** shows that the conventional regression model produced multiplicative effects confirming the results of the first (**Table 2**) (Boo et al.) and second case study (**Table 7**) (Boo et al. 2018b). Moreover, *Average Weight* and *Dogs per Capita* confirmed prior findings on the association between larger body sizes on higher canine cancer incidence (Eichelberg and Seine 1996; Michell 1999) and the relevance of the indicator dogs per capita as a proxy for urban status (Pospischil et al. 2013).

The power analysis of the conventional regression model returned a minimum sample size of 347 municipal units – a sample size that would guarantee acceptable statistical power (ß=.80), given a standard significance level (α=.05) and a small effect size (f^2=.04) (Ferguson 2009). The previous consideration was critical to the definition of regional scale models as these were fit to regions involving all possible centers and geographic scales spanning between 346 and 2,324 nearest municipal units. This produced a considerable number of regional models, summing up to 4,594,548, evaluated through robust summary statistics of the selected diagnostic measures.

The value-by-alpha map presented in **Figure 13A** denotes the presence of regional models with lower statistical performance in terms of $R^2_{McFadden}$ statistics. Their geographic distribution indicated regions where a finer specification of the baseline model would be necessary to better fit average canine cancer incidence rates (Cressie 1993; Lloyd 2014). As expected, regions of reduced model fit were determined in the mountainous regions and rural hinterland of the western part of the country, where lower average canine cancer incidence also occurred. These shared geographic configurations reflected that different levels of completeness in the SCCR data, associated with the underascertainment of cancer cases, were critical to the modeling effort (Bartlett et al. 2010; O'Neill et al. 2014). These regions also presented a large spread in the $R^2_{McFadden}$ measures, which was associated with smaller geographic scales as presented through the scalogram in **Figure 13B**. These critical effects of geographic scale suggested the importance of modeling average canine cancer incidence

rates locally, in contrast to more conventional global approaches (Lloyd 2010, 2014).

To support this finding, the value-by-alpha maps presented in **Figure 14** depict the effects of spatial non-stationarity for all the independent variables. The same model coefficients could even produce multiplicative effects contrasting prior research on canine cancer when estimated within different regions, particularly for *Average Age* (**Figure 14A**), *Females per Male* (**Figure 14B**), and *Average Weight* (**Figure 14C**). In one sense, these statistical associations could be linked to local selective underascertainment of cancer cases, as older dogs may be less likely to undergo regular veterinary check-ups (Bonnett and Egenvall 2010). Therefore, there are persisting negative multiplicative effects of *Average Age* in the regional models, the conventional regression model, and the models developed in the previous case studies (Boo et al. 2017, 2018b). However, it is also likely that known local and regional preferences concerning breeds could result in different multiplicative effects of *Average Age* and *Average Weight* across the study area (Pospischil et al. 2013).

Spatial non-stationarity was less striking for the confounding factors accounting for potential underascertainment of cancer cases. The reason was that these independent variables showed more robust statistical associations across the study area. Nevertheless, **Figure 15** demonstrates that all statistical associations in the model of average canine cancer incidence rates were impacted by effects of geographic scale to a certain extent. Again, critical effects were detected for *Average Age* (**Figure 15A**), *Females per Male* (**Figure 15B**), and *Average Weight* (**Figure 15C**). In summary, these results indicate that, when assessing average canine cancer incidence rates, effects of spatial non-stationarity and geographic scale complicate the interpretation of the statistical associations (Atkinson and Tate 2000; Tate and Atkinson 2001). However, the results also confirm that this issue can be successfully addressed through regional models because they account for the spatial structure within the statistical associations. As such, this modeling framework should also be considered for potential environmental-sentinel applications in Switzerland.

4.1.4. Environmental-sentinel applications

The ultimate aim of this thesis is to address the fourth knowledge gap (**KG 4**).

KG 4 For *potential environmental-sentinel applications*, it is crucial to carefully contextualize statistical inference based on the spatial epidemiology of canine cancer (Scotch et al. 2009). This issue involves an in-depth understanding of the generalizability of statistical associations estimated in models of canine cancer incidence.

To fill this particular knowledge gap, the three case studies examined canine cancer diagnostic cases retrieved from the SCCR – the largest and longest-lived canine cancer registry to date (Grüntzig et al. 2015). Similar to other existing canine cancer registries, such as the CANR and Tulsa Registry (Brønden et al. 2007), an essential characteristic of the SCCR is the standardized attribution of canine cancer diagnostic cases based on the ICD-O-3 classification method (Grüntzig et al. 2015, 2016). As discussed in extensive reviews on canine cancer registration by Brønden et al. (2007) as well as Nødtvedt et al. (2011), this standardization method is vital to compare evidence from different canine cancer registries.

Besides, as the ICD-O standardization method was originally conceived for classifying human tumors and cancers (WHO 2015), the SCCR also allows for comparative studies of canine and human cancers – a crucial stepping stone for potential environmental-sentinel applications (Rabinowitz et al. 2009; Scotch et al. 2009). The canine cancer diagnostic cases stored in the SCCR also include detailed information on the residential address of the diagnosed dogs (Boo et al. 2017, 2018a, b). This feature can enable the extension of the body of knowledge on canine cancer summarized, among others, by Schmidt (2009) and Reif (2011) by studying the geographic characteristic of the places where the disease occurs (Elliott et al. 1996).

In this regard, studying the SCCR employing the methods of spatial epidemiology could considerably update comparisons of the geographic distribution of canine and human cancers. These studies were mostly carried out in the 1960s and 1970s by Dorn et al. (1968a, b), Schneider et al. (1968), and Schneider (1970). To update the knowledge on the geographic

distribution of canine cancer frequencies, the three case studies developed a preliminary assessment based on disease mapping (Boo et al. 2017, 2018a, b). This study method is considered exceptional in the study of canine cancer frequencies (Kimura et al. 2015), particularly at the country level, and it was facilitated by the extraordinarily rich attribution of the SCCR data (Grüntzig et al. 2015).

The maps depicted three measures of the frequency of new canine cancer cases. **Figure 3** features crude canine cancer incidence across Swiss municipal units in 2008 (Boo et al. 2017). **Figure 11** depicts canine cancer incidence rates across Swiss municipal units for the same year (Boo et al. 2018b). Lastly, **Figure 12** shows average canine cancer incidence rates across Swiss municipal units for the period 2008-2013 (Boo et al. 2018a). Similar to the spatial epidemiology of human cancer, these measures referred to the new cases of canine cancer recorded within each municipal unit during a specific period (Lawson et al. 2001).

In spite of the use of different measures of canine cancer frequency, the three maps show very similar geographic distributions across the study area. These distributions are dominated by higher values in the eastern part of the country, typically within the major urban agglomerations, such as Zurich, Berne, and Basle. Such an arrangement implies that higher frequencies persist also when accounting for the geographic distribution of the at-risk dog population (i.e., by computing incidence rates) and the random variations associated with temporal variability (i.e., by computing average incidence rates) (Lawson et al. 2001; Lawson 2006).

These results contributed to speculation that the geographic distribution of canine cancer frequencies across Swiss municipal units results, at least in part, in different levels of completeness in the SCCR data. In this respect, taking into account that the data source was compiled by the only official laboratories for canine cancer diagnosis in the country (i.e., the Vetsuisse Institutes of Veterinary Pathology in Berne and Zurich), the underreporting of cancer cases was supposed to be negligible (Gibbons et al. 2014). Therefore, the different levels of completeness were considered to be associated with social and contextual factors resulting in the underascertainment of canine

cancer cases, for example, across urban and rural locations (Boo et al. 2017, 2018a, b).

This preliminary hypothesis on critical underascertainment of cancer cases was implicitly tested throughout the three case studies. For this purpose, the geographic-correlation study method was systematically employed (Elliott et al. 1996). The reason was that, by including demographic and confounding variables in the models of canine cancer frequency, they would describe the direction, strength, and significance level of the multiplicative effects associated with the different independent variables (Boo et al. 2017, 2018a, b). Owing to Heckman's seminal work on sample selection bias (Heckman 1974, 1976, 1979), these multiplicative effects uncovered the impact of potential confounding factors associated with the underascertainment of cancer cases.

In summary, the impact of the confounding variables within the different models suggested that underascertainment of cancer cases was critical to the statistical inference based on the models of canine cancer frequencies (Gibbons et al. 2014; O'Neill et al. 2014). Furthermore, statistical inference for potential environmental-sentinel applications was also shown to be affected by sample size and statistical power (first case study), as well as ecological fallacy associated with spatial data aggregation (second case study). However, by addressing these critical issues, we provided a first stepping stone towards the production of generalizable evidence for potential environmental-sentinel applications in Switzerland.

4.2. GENERAL LIMITATIONS

4.2.1. Epidemiologic considerations

The first set of limitations of the case studies presented in this thesis concerns the epidemiological nature of the investigation (Bartlett et al. 2010; O'Neill et al. 2014). In this regard, a crucial limitation was to consider cancer as a whole and assess the geographic distribution of canine cancer frequencies together for all cancer types stored in the SCCR (Boo et al. 2017, 2018a, b). Although this practice is not uncommon in the spatial epidemiology of human cancer (Roquette et al. 2017), existing comparative studies of canine and human cancers have mostly focused on specific cancer types (Schneider et al. 1968; O'Brien et al. 2000; Pastor et al. 2009). This was to investigate shared geographic distributions of cancer types and associated risk factors.

Considering cancer as a whole was required, on the one hand, by the fact that information on specific environmental exposures, such as ETS or air pollution in general, for specific cancer types was difficult or impossible to obtain retrospectively for the given study years. On the other hand, owing to the proposed research questions, concentrating on selected cancer types and relevant environmental exposures would have not necessarily provided more compelling results. The reason for that was the different level of underascertainment of cancer cases for various cancer types - some are more difficult to detect than others (Dorn et al. 1968b).

A second important limitation, which is also typical of the spatial epidemiology of human cancer (Boscoe et al. 2004), was related to the enumeration of canine cancer cases within municipal units on a yearly basis (Boo et al. 2017, 2018a, b). This practice is related to the assumption that the selected units of aggregation – the municipal unit and the year – provide a meaningful reflection of the epidemiological processes of interest (Ward and Wartenberg 2006). Furthermore, the assumption of the sedentary behavior within the unit of aggregation during the entire study period is essential (Elliott et al. 1996; Beale et al. 2008). Nevertheless, aggregating individual cancer cases over municipal units and years was meant to reduce spurious

correlations owing to sample variability (Lawson 2006). In addition, because of the relatively short life span of dogs, computing canine cancer frequencies on a yearly base was seen as a potential strategy to isolate risk factors and confounding variables within specific residential settings (Reif 2011).

4.2.2. Statistical modeling considerations

When assessing canine cancer frequencies, varying levels of completeness of the SCCR across the study area might have been an initial issue for the modeling efforts of the three case studies (Boo et al. 2017, 2018a, b). This limitation was partly corrected by implementing confounding variables associated with potential underascertainment of cancer cases (Elliott and Wakefield 2000; Rezaeian et al. 2007). However, the third case study showed that, within a variety of geographic regions, a finer model specification was needed to better assess canine cancer frequencies (Boo et al. 2017). This implies that the model could not accurately reflect the different levels of completeness in the data through the selected confounding variables (Bartlett et al. 2010; O'Neill et al. 2014). Yet, when considering canine cancer as a whole, instead of focusing on specific types and associated environmental exposures, might have been the basis for poor model specification (Reif 2011; Rowell et al. 2011).

The second modeling issue involves the selected modeling framework. This is because, in spite of significant overdispersion and zero-inflation in the measures of canine cancer frequencies, except for the second case study, no alternative to the Poisson models was tested. Still, for example, overdispersion might have produced dramatically inflated confidence intervals for the coefficient estimated through the Poisson models (Cameron and Trivedi 1990; Berk and MacDonald 2008). In spite of such a potential limitation, this choice was because of the comparative frameworks developed in the two case studies, namely the complete enumeration versus the filtered sample (Boo et al. 2018b) and the classical regression model versus the regional models (Boo et al. 2018a). This was because the pairwise comparison of models involving additional parameters accounting for overdispersion and zero inflation would have challenged the interpretation of the individual multiplicative effects (Arab 2015).

The last modeling limitation concerns the geographic distribution of the underlying canine cancer incidence examined through the different measures of canine cancer frequencies. Considering that these persistently exhibited strong geographic configurations, spatial autocorrelation in the data could violate the assumption of independence of observations in the different models (Wall 2004). This condition of spatial autocorrelation might have been critical to the statistical performance of the models of canine cancer frequencies. This is because spatial autocorrelation must be addressed through particular modeling frameworks, such as the SAR (Whittle 1954) and CAR (Besag 1974) models. However, these models have not been implemented because they require a deeper comprehension of the correlation structures within the data (Wall 2004).

Given the contribution of the three case studies to the knowledge gaps identified in Section 2.4, the next section provides an answer to the specific research questions.

4.3. REVISITING THE RESEARCH QUESTIONS

4.3.1. Society and context

Hereafter, the general discussion presented in Section 4.1 is reviewed to answer the first research question (RQ 1). In detail:

RQ 1 How do *society and context* challenge the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This research question was implicitly addressed throughout the different models of canine cancer frequencies examined in the three case studies (Boo et al. 2017, 2018a, b). For this purpose, the multiplicative effects associated with demographic risk factors and, particularly, potential confounding factors accounting for the underascertainment of canine cancer cases were systematically examined. The reason for that was testing the significance, direction, and strength of the statistical associations for providing a better understanding of societal and contextual influences on the ascertainment of canine cancer cases and consequent concerns on data quality.

In this regard, the case studies indicated that urban character, socio-economic status, and distance to veterinary care were significant confounding variables, associated with societal and contextual effects on the owner decisions surrounding veterinary care. These confounding factors suggested that specific societal and contextual settings (i.e., rural, less wealthy, and distant from veterinary care municipalities) systematically relate with negative multiplicative effects on the geographic distribution of canine cancer frequencies. This finding provided evidence for arguing that society and context could be held, at least partially, accountable for the underascertainment of cancer cases. The consequent issues of data quality could be dealt with by factoring in the confounding factors within the model of canine cancer incidence.

The first case study also exhibited the extreme effects of underascertainment of cancer cases by investigating the impacts of structural

zeros on models of canine cancer incidence. Through model cross-validation, structural zeros were found to be critical both to the statistical performance and predictive power of the models. The results of this assessment demonstrated that the societal and contextual settings described before might challenge the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors, and, ultimately, the production of generalizable evidence. However, identifying these issues of data quality enable addressing them through a finer model specification.

4.3.2. Spatial data

Hereafter, the general discussion presented in Section 4.1 is reviewed to answer the second research question (RQ 2). In detail:

RQ 2 What are the implications of using *spatial data* in the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This research question was addressed in the second case study (Boo et al. 2018b), which evaluated the effects of spatial data aggregation on the statistical performance of models of canine cancer incidence rates. Such an assessment was developed by contrasting two modeling scenarios — based on standard municipal units and dasymetrically refined units, defined as the part of residential land within the municipal unit. These two types of enumeration units were employed for aggregating canine cancer incidence and independent variables. Still, changes were only induced in the independent variables based on the spatial extent relative location of the enumeration unit.

This comparative assessment demonstrated that, when embedded in the model of canine cancer incidence rates, the density variable was statistically significant only when employing dasymetrically refined units. The reason was that dasymetric refinement enabled computing more meaningful geographic distributions of the indicator considering geographic context. However, also because of the specific character of the study area, the distance variable did not perform differently between the two enumeration

units. These contrasting results indicated that effects of spatial data aggregation are difficult to fully mitigate.

The second case study revealed that spatial data aggregation impacts the estimation of the statistical associations between the geographic distribution of canine cancer incidence rates and associated risk factors. Still, the improvement in the statistical performance of the model of canine cancer incidence rates based on dasymetrically refined units reflects that this method can deal with the effects of spatial aggregation because dasymetric refinement facilitates representing the geographical distribution of the explanatory variables more accurately.

4.3.3. Analytical framework

Hereafter, the general discussion presented in Section 4.1 is reviewed to answer the third research question (RQ 3). In detail:

RQ 3 How does the selected *analytical framework* impact the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors?

This research question was addressed in the second and third case study (Boo et al. 2018b, c), specifically by assessing the limitations of the conventional regression model in estimating statistical associations. For this purpose, the proposed exploratory framework was developed based on the concept of the regional model. Regional models were fit across the municipal units within the study area at different geographic scales. The resulting model diagnostics were lastly contrasted with the one of a standard regression model. Such a comparative framework was utilized to assess the effects of spatial non-stationarity and geographic scale in a model of average canine cancer incidence rates.

This comprehensive assessment enabled detecting remarkable differences in the statistical performance of regional models across both the study area and geographic scales (Leyk et al. 2012b; Maclaurin et al. 2015). This finding suggested that potential issues, such as model misspecification and different levels of completeness in the SCCR data, could be crucial in certain regions within the study area (Brunsdon et al. 1996). As a

consequence, the conventional regression model would poorly reflect the underlying processes associated with the geographic distribution of the average canine cancer incidence rates.

The effects of spatial non-stationarity and geographic scale produced substantial variations in the multiplicative effects associated with the different coefficient estimates. Not surprisingly, these variations were considerably larger at small geographic scales. Such a result further indicated that geographical structure is vital for modeling average canine cancer incidence rates. To address this issue, the conventional regression model should be replaced by more local or regional modeling approaches in the estimation of statistical associations between the geographic distribution of canine cancer frequencies and associated risk factors.

4.3.4. Statistical inference

Hereafter, the general discussion presented in Section 4.1 is reviewed to answer the fourth research question (**RQ 4**). In detail:

RQ 4 How does the estimation of statistical associations between the geographic distribution of canine cancer and associated risk factors impact *statistical inference* for potential environmental-sentinel applications?

The fourth research question was implicitly investigated throughout the three case studies (Boo et al. 2017, 2018a, b). In particular, the issue of generalizable evidence was dealt with in the first case study (Boo et al. 2018b) by assessing the predictive power of a model of canine cancer incidence. Furthermore, potential ecological fallacies associated with spatial data aggregation were tackled in the second case study (Boo et al. 2018a) by comparing two types of enumeration units, namely municipal units and dasymetrically refined units.

The first case study demonstrated that structural zeros, an extreme manifestation of the underascertainment of cancer cases, could be vital to generalizability. Furthermore, known considerations about sample size and resulting statistical power were found to challenge the ability to successfully identify an effect. For this reason, it is crucial to consider sample size for

statistical inference based on statistical associations between the geographic distribution of canine cancer and associated risk factors.

The second case study indicated that spatial data aggregation could involve ecological fallacies when interpreting the model of canine cancer incidence rates. This matter is especially challenging when statistical inference based on models of canine cancer incidence rates are regarded as directly generalizable. As expected, the impacts of the MAUP, and, more generally, changes of support problems were found to be critical when computing variables embedded in the model of canine cancer incidence as the latter vary as a function of the shape and areal extent of the enumeration unit. The unit of spatial data aggregation should be accounted for with care in potential environmental-sentinel applications.

As anticipated in Section 1.2.2, this thesis is grounded in the body of knowledge of GIScience. These connections will be examined hereafter to foster interdisciplinary research across GIScience and the spatial epidemiology of canine cancer.

4.4. CONNECTIONS TO GISCIENCE

This thesis is a stepping stone towards comparative studies of canine and human cancers in Switzerland. However, given that just a few comparative studies have adopted the methods of spatial epidemiology to date, it is essential to examine how this thesis melds with the discipline of GIScience. These connections are described hereafter by referring to the thematic areas of the body of knowledge of GIScience presented in Section 1.2.2.

Analytics and modeling

This thesis involves a set of connections to GIScience in the thematic knowledge area "Analytics and Modeling." This knowledge area is broadly connected with the creation of knowledge surrounding geographically explicit processes and their distributions (DiBiase et al. 2017). Specific topics of contribution are summarized in the latest version of the "Geographic Information Science & Technology Body of Knowledge" (DiBiase et al. 2017), providing a basic reference to ground this thesis in the discipline of GIScience. The specific topics of contributions to "Analytics and Modeling" (**AM**) and related research questions (**RQ**) are reported in parentheses.

A crucial element grounding this thesis into GIScience was the systematic use of E(S)DA methods (**AM 19**), namely the different maps of the geographic distribution of canine cancer frequencies. Examining these cartographic representations was paramount for developing a basic understanding of the spatial and statistical associations within the study area. This led, for instance, to a better specification of the confounding variables implemented in the different models of canine cancer frequencies. Most of these models were built based on a Poisson regression framework, one of the most common statistical methods of spatial econometrics (**AM 31**), which is also commonly employed for assessing incidence and rates of rare diseases, such as cancer.

Furthermore, the Poisson regression framework was applied in the development of original analytical procedures to investigate the proposed research questions (**AM 53**). First, model cross-validation enabled detecting the effects of structural zeros on models of canine cancer incidence, suggesting

critical issues of spatial data quality (**RQ 1 – AM 49**). Second, assessing the impacts of spatial data aggregation allowed for a better understanding of challenges in the computation of the variables included in models of canine cancer incidence rates (**QR 2 – AM 50**). Lastly, modeling average canine cancer incidence rates within different regions of the study area exposed the effects of spatial non-stationarity and geographic scale (**RQ 3 - AM 34**).

Cartography and visualization

This thesis also connects to the thematic knowledge area "Cartography and Visualization." This important knowledge area of GIScience involves concerns the general design and use of maps and mapping technology (DiBiase et al. 2017). Specific topics of contribution are also reported in the latest version of the "Geographic Information Science & Technology Body of Knowledge" (DiBiase et al. 2017), again providing a basic reference to ground this thesis in the discipline of GIScience. The specific topics of contributions to "Cartography and Visualization" (**CV**) and related research questions (**RQ**) are once more reported in parentheses.

A primary connection to GIScience was in designing thematic maps of the geographic distribution of canine cancer frequencies and other modeling features for visual analytics purposes (**RQ 1 – CV 23**). Furthermore, design considerations concerning choropleth and dasymetric maps of canine cancer frequencies were discussed to evaluate the strengths and limitations of the two cartographic techniques (**CV 11**). Contrasting the geographic distributions depicted with choropleth and dasymetric maps also enabled interpretation of the effects of spatial data aggregation associated with changes of support problems (**RQ 2 – CV 22**). These problems were further discerned in the model of canine cancer incidence rates presented in the previous section.

Another relevant connection was in the design and interpretation of value-by-alpha maps, a particular type of bivariate cartographic technique (**CV 22**), recently proposed as an alternative to the cartogram (**CV 32**). Value-by-alpha maps were employed to map uncertainty in the statistical performance and multiplicative effects within a conventional regression model of average canine cancer incidence rates (**CV 18**). Besides this, combining value-by-alpha maps with an additional visualization technique, in particular the scalogram,

granted crucial insights into the effects of spatial non-stationarity and geographic scale in the model of average canine cancer incidence rates (**RQ 3 – CV 36**).

GIScience & technology and society

Lastly, this thesis connects to the thematic knowledge area, "GIScience & Technology and Society." This area consists of considering the different impacts of GIScience, from the institution down to the individual level (DiBiase et al. 2017). The specific topics of contribution to this knowledge area are reported in the latest version of the "Geographic Information Science & Technology Body of Knowledge" (DiBiase et al. 2017), which once more provides a basic reference to ground this thesis in the discipline of GIScience. Again, the specific topics of contributions to "GIScience & Technology and Society" (**GS**) and related research questions (**RQ**) are reported in parentheses.

As highlighted by the topic of aggregation of spatial entities (**GS 20**), the most important link to GIScience consisted of raising awareness about the geographic nature of the spatial epidemiology of canine cancer, in other words, in further endorsing the motto, "spatial is special" (Goodchild 1992). The reason for that was to develop a better understanding of specific challenges and limitations associated with society and context (**RQ 1**), spatial data (**RQ 2**), and analytical framework (**RQ 3**). Answering these research questions enabled demonstration of several crucial connections between the spatial epidemiology of canine cancer and selected thematic knowledge areas of GIScience.

The last topic grounding this thesis in the discipline of GIScience is the ubiquitous subject of citizen science (**GS 24**). As previously mentioned, one of the leading challenges of the spatial epidemiology of canine cancer is spatial data quality (**AM 49**). This is because of different levels of completeness based on the underascertainment of cancer cases. It is hence vital to consider that behind every ascertained case there is a dog owner that acts as a sort of non-professional volunteer in a scientific endeavor. Identifying challenges and limitations associated with society and context (**RQ 1**) and statistical inference (**RQ 4**) for sentinel purposes was, therefore, the first step towards better inclusion of non-scientists in the spatial epidemiology of canine cancer.

The next chapter summarizes the achievements of this thesis and provides an outlook on directions for future work extending the connections between GIScience and the spatial epidemiology of canine cancer.

5. CONCLUSIONS AND OUTLOOK

Owing to the proposed research goals, the first section of this chapter summarizes the achievements of this thesis. These connect with the objective of addressing the challenges and limitations of the spatial epidemiology of canine cancer for potential environmental-sentinel applications. Identifying the research achievements enable the proposition of an outlook on directions for future work, concerning both comparative studies of canine and human cancers in Switzerland and analytical refinements.

5.1. INSIGHTS AND ACHIEVEMENTS

This thesis allowed for the understanding and addressing of challenges and limitations of the spatial epidemiology of canine cancer for potential environmental-sentinel applications. The achievements are directly connected with the four groups of challenges and limitations of the spatial epidemiology of human cancer highlighted by Jacquez (2004). In detail:

1) Society and context

The first case study emphasized that the spatial epidemiology of canine cancer is concerned with challenges and limitations associated with the presence of structural zeros. These occur when zero incidence originates from the sole absence of performed diagnostic examinations within a given sample unit because of the underascertainment of cancer cases. Societal and contextual causes of structural zeros and the resulting impacts on models of canine cancer incidence were thoroughly investigated. This was through examining the multiplicative effects of demographic and confounding variables accounting for potential under-ascertainment of cancer cases.

Such an analytical framework facilitated identification of the critical influences of society and context, such as urban character, socioeconomic status, and distance to veterinary care. However, through model cross-validation, the impacts of structural zeros were identified in an overall

decrease of the statistical performance and predictive power of the model of canine cancer incidence. These critical findings suggested potential causes of structural zeros and their effects on the estimation of statistical associations between the geographic distribution of canine cancer incidence and associated risk factors.

This permitted contending that modeling the geographic distribution of canine cancer frequencies necessitates the inclusion of explanatory variables accounting for societal and contextual settings. This is because these variables enable, on the one hand, better understanding whether the geographic distributions are influenced by phenomena that are not risk factors but confounders associated with selection bias. On the other hand, including these variables allows finer insights into geographic distributions because adjusting canine cancer frequencies for these confounders fosters a more accurate assessment of the potential risk factors.

2) Spatial data

The second case study demonstrated that the spatial epidemiology of canine cancer is affected by challenges and limitations resulting from the process of spatial data aggregation. The reason is that assessing enumerated canine cancer cases in the form of incidence or rates implies that summary statistics and statistical associations to the independent variables can change according to the shape and areal extent of the enumeration unit. This issue was investigated by systematically contrasting two scenarios for modeling canine cancer incidence rates. The first scenario made use of municipal units and the second dasymetrically refined units, which are defined as the portion of residential land within the municipal unit.

This comparative investigation uncovered critical effects of spatial data aggregation, especially when computing the density variable (i.e., human population density) included in the model of canine cancer incidence rates. Furthermore, the two scenarios produced models of canine cancer incidence rates featuring different statistical performance, with a better goodness-of-fit for the model using dasymetrically refined units. These finding demonstrated that spatial data aggregation impacts the estimation of statistical associations between the geographic distribution of canine cancer incidence rates and

associated risk factors. Moreover, dasymetric refinement was successful in, at least partly, mitigating such effects.

When modeling the geographic distribution of canine cancer frequencies, it is imperative to consider the results at different levels of aggregation and for various types of spatial units. This is because this type of geographic analysis is always affected by the effects of spatial data aggregation. In this regard, a comparative approach will enable choosing both explanatory variables that are less affected by spatial data aggregation and the level of aggregation and type of spatial unit that produces more meaningful explanatory variables.

3) Analytical framework

The third case study emphasized that effects of spatial non-stationarity and geographic scale challenge the use of classical analytical frameworks employed in the spatial epidemiology of canine cancer. For this purpose, models of average canine cancer incidence rates where fit across multiple regions within the study area for a range of geographic scales. Diagnostic summaries across the regional models were then computed, summarized and mapped. Lastly, the statistical performance and associations were contrasted with the same diagnostics for a conventional regression model.

Such a comparative evaluated enabled identifying remarkable variations in the goodness-of-fit across both the study area and geographic scales. Furthermore, statistical associations related to different coefficient estimates featured critical effects of spatial non-stationary and geographic scale. These analytical issues were mostly detected for the regional models at small geographic scales. With this, these results make a case for the use of local or regional models for the estimation of statistical associations between the geographic distribution of average canine cancer incidence rates and associated risk factors.

This is because modeling the geographic distribution of canine cancer frequencies requires analytical frameworks that accommodate the characteristics of spatial phenomena. These are, among others, related with spatial variations in the statistical associations or, in other words, spatial non-stationarity, as well as the spatial extent or geographic scale of the study

area under consideration. When failing to address these conditions, the selected analytical framework only accounts for relationships between canine cancer frequencies and explanatory variables in an a-spatial context by ignoring the impact of spatial structure and interaction among analytical units.

4) Statistical inference

The three case studies showed that the previously discussed challenges and limitations of the spatial epidemiology of canine cancer also influenced statistical inference for potential environmental-sentinel applications. In particular, the first case study showed that the statistical performance and predictive power of the model of canine cancer incidence were challenged by sample size and statistical power. These elements were found to be critical for producing generalizable evidence, thus suggesting extreme caution in considering statistical inference for potential environmental-sentinel applications.

Moreover, the second case study shed light on the critical impacts of spatial data aggregation on statistical inference for potential environmental-sentinel applications. The reason for that is ecological fallacy, in its geographical manifestation of the MAUP, as statistical inference based on spatial units should not transfer to the individual level. For the determination of statistical inference, the third case study also showed that the model of average canine cancer incidence rates was challenged by effects of spatial non-stationary and geographic scale. Utilizing an analytical framework accounting for these issues is, therefore, vital to producing inference for potential environmental-sentinel applications.

By addressing these general challenges and limitations of the spatial epidemiology of canine cancer, this thesis offers a stepping stone towards environmental-sentinel applications. For this reason, future work will involve research into two directions – developing comparative studies of canine and human cancers in Switzerland, and extending the knowledge of the phenomena examined in the three cases studies found in this thesis.

5.2. OUTLOOK AND FUTURE WORK

Developing comparative studies in Switzerland

This thesis was motivated by potential environmental-sentinel applications enabled by the spatial epidemiology of canine cancer. To assess challenges and limitation affecting this particular study approach, we examined the geographic distribution of canine cancer frequencies retrieved from the SCCR – the largest and longest-lived canine cancer registry to date. Given the exceptional character of this data source, a natural step forward towards environmental-sentinel applications will be developing comparative studies of canine and human cancers in the Swiss context (Grüntzig et al. 2015, 2016). To this end, the geographic distribution of canine cancer frequencies ought to be assessed against human cancer frequencies.

In this regard, the Swiss National Institute for Cancer Epidemiology and Registration (NICER) will play a major role. Founded in 2007, NICER compiles and aggregates human cancer records collected by the cantonal and regional cancer registries of Switzerland (NICER 2017). However, owing to privacy considerations, only a small number of attributes is currently available to research. Furthermore, the residential location is only provided at a coarse administrative level, such as the Canton (NICER 2017). This known limitation of spatial epidemiology requires addressing by contacting the individual cancer registries as these can influence privacy and access regulations.

Comparative studies of canine and human cancers in Switzerland will drive future research of the SCCR towards two directions. The first will be further exploring the promising venue of developing sentinel applications at the cellular and individual level (Pospischil et al. 2015). The second direction will be better understanding the SCCR data and tailoring specific study methods at the ecologic or population level. This is because the challenges and limitations highlighted in this thesis confirm the need for more accurate assessments of shared geographic distribution of canine and human cancer incidence to propose effective environmental sentinel applications.

Extending the three case studies

Further dealing with the challenges and limitations of spatial epidemiology will necessitate extending the conceptual and methodological connections to GIScience. Additional efforts will be carried out in considering the impacts of society and context. As such, a finer specification of the confounding variables will be vital to developing comparative studies (Heckman 1979). Additionally, comparing human and canine cancer frequencies within specific societal and contextual settings influencing the use of veterinary care, such as remote, rural, and economically deprived regions of Switzerland, might inevitably lead to uncertain statistical associations (Elliott and Wartenberg 2004). This issue will require delineating different regimes of underascertainment of cancer cases for selecting regions of reduced incompleteness in the SCCR (Boo et al. 2016).

A simple binary dasymetric refinement of Swiss municipal units to their portion of residential land reduced the effects of spatial data aggregation to some extent. To further tackle this issue, more complex dasymetric refinement techniques are necessary (Eicher and Brewer 2001). These will, for example, involve defining additional ancillary variables to better approximate the geographic distribution of the at-risk populations across residential land within the municipal units (Mennis 2009). A last direction for future work will be better model specification by both including relevant risk factors and considering modeling frameworks that accommodate more effectively the spatial (i.e., spatial autocorrelation (Lloyd 2010; Wall 2004)) and statistical (i.e., overdispersion and/or zero inflation (Zeileis et al. 2008)) distribution of the incidence retrieved from the SCCR.

In conclusion, this thesis addressed a set of important challenges and limitations related to the spatial epidemiology of canine cancer. The results underscored several connections between spatial epidemiology and GIScience. Owing to the crucial impact of potential environmental-sentinel applications, further interdisciplinary research across these disciplines is advocated.

6. REFERENCES

Ahrens W, Krickeberg K, Pigeot I (2005) An Introduction to Epidemiology. In: Ahrens W, Pigeot I (eds) Handbook of Epidemiology. Springer, Berlin, Germany, pp 1–40

Akaike H (1974) A New Look at the Statistical Model Identification. IEEE Trans Autom Control 19:716–723

Al-Ahmadi K, Al-Zahrani A (2013) Spatial Autocorrelation of Cancer Incidence in Saudi Arabia. Int J Environ Res Public Health 10:7207–7228

Alexander FE, Boyle P (eds) (1997) Methods for Investigating Localized Clustering of Disease. IARC Sci Publ 1–20

ANIS (2017) Animal Identity Service AG. http://www.anis.ch. Accessed 31 Dec 2017

Anselin L (1995) Local Indicators of Spatial Association — LISA. Geogr Anal 27:93–115

—— (1988) Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht, The Netherlands

Apperly FL (1941) The Relation of Solar Radiation to Cancer Mortality in North America. Cancer Res 1:191–195

Arab A (2015) Spatial and Spatio-Temporal Models for Modeling Epidemiological Data with Excess Zeros. Int J Environ Res Public Health 12:10536–10548

Atkinson PM, Tate NJ (2000) Spatial Scale Problems and Geostatistical Solutions: A Review. Prof Geogr 52:607–623

Bailey TC, Gatrell AC (1995) Interactive Spatial Data Analysis. Addison Wesley Longman, Harlow, UK

Baioni E, Scanziani E, Vincenti MC, et al (2017) Estimating canine cancer incidence: findings from a population-based tumour registry in northwestern Italy. BMC Vet Res 13:203

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. CRC Press, Boca Raton, FL, US

Bartlett PC, Van Buren JW, Neterer M, Zhou C (2010) Disease surveillance and referral bias in the veterinary medical database. Prev Vet Med 94:264–271

Basseville M, Benveniste A, Willsky AS (1992) Multiscale Autoregressive Processes. IEEE Trans Signal Process 40:1915–1934

Bavaud F (1998) Models for Spatial Weights: A Systematic Look. Geogr Anal 30:153–171

Beale L, Abellan JJ, Hodgson S, Jarup L (2008) Methodologic Issues and Approaches to Spatial Epidemiology. Environ Health Perspect 116:1105–1110

Beam C (2013) Biostatistical Applications in Cancer Research. Springer, New York, NY, US

Berger VW, Zhang J (2005) Structural Zeros. In: Everitt BS, Howell DC (eds) Encyclopedia of Statistics in Behavioral Science. John Wiley, Chichester, UK

Berk R, MacDonald JM (2008) Overdispersion and Poisson Regression. J Quant Criminol 24:269–284

Besag J (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. J R Stat Soc Series B Stat Methodol 36:192–236

Best N, Wakefield J (1999) Accounting for inaccuracies in population counts and case registration in cancer mapping studies. J R Stat Soc Ser A Stat Soc 162:363–382

Bivand RS (1998) A review of spatial statistical techniques for location studies. University of Bergen, Department of Geography, Bergen, Norway

Bliss RL, Katz JN, Wright EA, Losina E (2012) Estimating proximity to care: Are straight line and zipcode centroid distances acceptable proxy measures? Med Care 50:99–106

Blot W, Fraumeni J (1982) Geographic epidemiology of cancer in the United States. In: Schottenfeld D, Fraumeni FJ (eds) Cancer Epidemiology and Prevention. WB Saunders, New York, NY, US, pp 179–193

——— (1977) Geographic patterns of oral cancer in the United States: Etiologic implications. J Chronic Dis Manag 30:745–757

Bonnett BN, Egenvall A (2010) Age patterns of disease and death in insured Swedish dogs, cats and horses. J Comp Pathol 142 Suppl 1:33–38

Boo G, Fabrikant SI, Leyk S (2015) A novel approach to veterinary spatial epidemiology: dasymetric refinement of the Swiss Dog Tumor Registry data. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci II-3/W5:263–269

Boo G, Leyk S, Fabrikant SI, et al (2018b) Exploring uncertainty in canine cancer data sources through dasymetric refinement. ISPRS Int J Geo-Inf [submitted]

Boo G, Leyk S, Brunsdon C, et al (2018a) The importance of regional models in assessing canine cancer incidences in Switzerland. PLOS ONE.

Boo G, Leyk S, Fabrikant SI, et al (2017) Assessing effects of structural zeros on models of canine cancer incidence: a case study of the Swiss Canine Cancer Registry. Geospat Health 12:121–129

Boo G, Leyk S, Fabrikant SI, Pospischil A (2016) A regional approach for modeling dog cancer incidences with regard to different reporting practices. In: Miller JA, O'Sullivan D, Wiegand N (eds) Ninth International Conference on GIScience Short Paper Proceedings. Springer, Heidelberg, Germany, pp 29–32

Boscoe FP, Ward MH, Reynolds P (2004) Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. Int J Health Geogr 3:28

Boyle P, Muir CS, Grundmann E (2012) Cancer Mapping. Springer, Berlin, Germany

Breslow NE, Enstrom JE (1974) Geographic correlations between cancer mortality rates and alcohol-tobacco consumption in the United States. J Nat Cancer Inst 53:631–639

Brønden LB, Flagstad A, Kristensen AT (2007) Veterinary Cancer Registries in Companion Animal Cancer: A Review. Vet Comp Oncol 5:133–144

Brønden LB, Nielsen SS, Toft N, Kristensen AT (2010) Data from the Danish veterinary cancer registry on the occurrence and distribution of neoplasms in dogs in Denmark. Vet Rec 166:586–590

Bronson RT (1982) Variation in age at death of dogs of different sexes and breeds. Am J Vet Res 43:2057–2059

Brown ML, Potosky AL, Thompson GB, Kessler LK (1990) The knowledge and use of screening tests for colorectal and prostate cancer: data from the 1987 National Health Interview Survey. Prev Med 19:562–574

Browne MW (2000) Cross-Validation Methods. J Math Psychol 44:108–132

Brunsdon C, Fotheringham AS, Charlton ME (2002) Geographically weighted summary statistics — a framework for localised exploratory data analysis. Comput Environ Urban Syst 26:501–524

——— (1996) Geographically weighted regression: A method for exploring spatial non-stationarity. Geogr Anal 28:281–298

Bukowski JA, Wartenberg D (1997) An alternative approach for investigating the carcinogenicity of indoor air pollution: Pets as sentinels of environmental cancer risk. Environ Health Perspect 105:1312–1319

Bukowski JA, Wartenberg D, Goldschmidt M (1998) Environmental causes for sinonasal cancers in pet dogs, and their usefulness as sentinels of indoor cancer risk. J Toxicol Envirol Health 54:579–591

Burbank F (1971) Patterns in Cancer Mortality in the United States 1950–1967, National Cancer Institute Monograph. US Government Printing Office, Washington DC, US

Burnham KP, Anderson D (2003) Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York, NY, US

Cameron CA, Trivedi PK (1990) Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. J Appl Econom 1:29–53

——— (1986) Regression-Based Tests for Overdispersion in the Poisson Models. J Econom 46:347–364

Cameron CA, Windmeijer FAG (1997) An R-squared measure of goodness of fit for some common nonlinear regression models. J Econom 77:329–342

——— (1996) R-Squared measures for count data regression models with applications to health-care utilization. J Bus Econ Stat 14:209–220

Carpenter LM, Beresford S a. A (1986) Cancer mortality and type of water source: Findings from a study in the UK. Int J Epidemiol 15:312–319

Cattin P (1980) Estimation of the Predictive Power of a Regression Model. J Appl Psychol 65:407–414

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250

Chen H, Hailey D, Wang N, Yu P (2014) A review of data quality assessment methods for public health information systems. Int J Environ Res Public Health 11:5170–5207

Cheng EM, Atkinson PM, Shahani AK (2011) Elucidating the spatially varying relation between cervical cancer and socio-economic conditions in England. Int J Health Geogr 10:51

Cho S-H, Lambert DM, Chen Z (2010) Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. Appl Econ Lett 17:767–772

Choi KM, Serre ML, Christakos G (2003) Efficient mapping of California mortality fields at different spatial scales. J Expo Sci Environ Epidemiol 13:120–133

Chou KC, Willsky AS, Benveniste A (1994) Multiscale recursive estimation, data fusion, and regularization. IEEE Trans Autom Control 39:464–478

Clark K (1997) Getting Started with Geographic Information System. Prentice Hall, Upple Saddle River, NJ, US

Cleek RK (1979) Cancers and the Environment: The Effect of Scale. Soc Sci Med [Med Geogr] 13:241–247

Clegg LX, Reichman ME, Miller BA, et al (2009) Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results: National Longitudinal Mortality Study. CCC 20:417–435

Cliff AD, Ord JK (1973) Spatial autocorrelation, monographs in spatial environmental systems analysis. Pion, London, UK

Cohen J (1995) The Earth is Round (p<. 05): Rejoinder. Am Psychol 50:997–1003

——— (1992) A Power Primer. Psychol Bull 112:155–159

——— (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Routledge, Hillsdale, NJ, US

Cressie NA (1996) Change of Support and the Modifiable Areal Unit Problem. J Geograph Syst 3:159–180

——— (1993) Inference for Lattice Models. In: Cressie NA (ed) Statistics for Spatial Data, 2nd edn. Wiley, Hoboken, NJ, US, pp 453–572

Cressie NA, Calder CA, Clark JS, et al (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–570

Cruickshank DB (1947) Regional Influences in Cancer. Br J Cancer 1:109–128

——— (1940) The Topography of the Relative Distribution of Cancer and Tuberculosis. Tubercle 21:281–291

d'Onofrio A, Mazzetta C, Robertson C, et al (2016) Maps and atlases of cancer mortality: A review of a useful tool to trigger new questions. Ecancermedicalscience 10:670

De Vita VT, Rosenberg SA (2012) Two Hundred Years of Cancer Research. N Engl J Med 366:2207–2214

Delamater PL, Messina JP, Shortridge AM, Grady SC (2012) Measuring geographic access to health care: raster and network-based methods. Int J Health Geogr 11:

DiBiase D, DeMers M, Johnson A, et al (2017) Geographic Information Science and Technology Body of Knowledge. http://gistbok.ucgis.org. Accessed 31 Dec 2017

——— (2006) Geographic Information Science and Technology Body of Knowledge. Association of American Geographers, Washington, DC, US

Dobson JM (2013) Breed-Predispositions to Cancer in Pedigree Dogs. ISRN Vet Sci 2013:1–20

Dobson JM, Samuel S, Milstein H, et al (2002) Canine neoplasia in the UK: Estimates of incidence rates from a population of insured dogs. J Small Anim Pract 43:240–246

Domencich TA, McFadden D (1975) Statistical Estimation of Choice Probability Functions. In: Urban Travel Demand — A Behavioral Analysis. North-Holland Publishing Co, New York, NY, US, pp 101–125

Dorn CR (1967) The Epidemiology of Cancer in Animals. Calif Med 107:481–489

Dorn CR, Taylor DON, Schneider R, et al (1968a) Survey of animal neoplasms in Alameda and Contra Costa Counties, California. II. Cancer morbidity in dogs and cats from Alameda County. J Nat Cancer Inst 40:307–318

Dorn CR, Taylor DON, Frye FL, Hibbard HH (1968b) Survey of animal neoplasms in Alameda and Contra Costa counties, California. I. Methodology and description of cases. J Nat Cancer Inst 40:295–305

Dvorzak M, Wagner H (2016) Sparse Bayesian Modelling of Underreported Count Data. Stat Modelling 16:24–46

Dykes J, Brunsdon C (2007) Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis. IEEE Trans Vis Comput Graphics 13:1161–1168

Edling C, Comba P, Axelson O, Flodin U (1982) Effects of low-dose radiation — A correlation study. Scand J Work Environ Health 8:59–64

Eichelberg H, Seine R (1996) Life expectancy and cause of death in dogs — The situation in mixed breeds and various dog breeds. Berl Munch Tierarztl Wochenschr 109:292–303

Eicher CL, Brewer CA (2001) Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. Cartogr Geogr Inf Sci 28:125–138

Elliott P, Cuzick J, English D, et al (1996) Geographical Epidemiology and Ecological Studies. In: Elliott P, Cuzick J (eds) Geographical and Environmental Epidemiology — Methods for Small-Area Studies. Oxford University Press, Oxford, UK, pp 10–24

Elliott P, Wakefield J (2000) Bias and confounding in spatial epidemiology. In: Elliott P, Wakefield J, Best N, Briggs D (eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, UK, pp 68–84

Elliott P, Wartenberg D (2004) Spatial Epidemiology: Current Approaches and Future Challenges. Environ Health Perspect 112:998–1006

Erath A, Löchl M, Axhausen KW (2008) Graph-theoretical analysis of the Swiss road and railway networks over time. Netw Spat Econ 9:379–400

FCI (2017) FCI Breeds Nomenclature. http://www.fci.be/en/nomenclature. Accessed 31 Dec 2017

Ferguson CJ (2009) An effect size primer: A guide for clinicians and researchers. Prof Psychol Res Pr 40:532–538

Fiscella K, Franks P, Gold MR, Clancy CM (2000) Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care. JAMA 283:2579–2584

FOPH (2017) Federal Office of Public Health - MedReg. http://www.medregom.admin.ch. Accessed 31 Dec 2017

Foster SA, Gorr WL (1986) An adaptive filter for estimating spatially-varying parameters: Application to modeling police hours spent in response to calls for service. Manag Sci 32:878–889

Fotheringham AS (1989) Scale-Independent Spatial Analysis. In: Goodchild MF, Gopal S (eds) The Accuracy Of Spatial Databases. CRC Press, Boca Raton, FL, US, pp 144–148

Fotheringham AS, Brunsdon C (1999) Local Forms of Spatial Analysis. Geogr Anal 31:340–358

Fotheringham AS, Brunsdon C, Charlton ME (2003) Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley, Chichester, UK

Fotheringham AS, Charlton ME, Brunsdon C (1996) The geography of parameter space: An investigation of spatial non-stationarity. Geog Inf Sys 10:605–627

Fotheringham AS, Wong DWS (1991) The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. Environ Plann A 23:1025–1044

Fox J (2015) Collinearity and Its Purported Remedies. In: Applied Regression Analysis and Generalized Linear Models, 3rd edn. Sage Publications, Thousand Oaks, CA, US, pp 307–331

Frome EL (1983) The Analysis of Rates Using Poisson Regression Models. Biometrics 39:665–674

Frome EL, Checkoway H (1985) Use of Poisson Regression Models in Estimating Incidence Rates and Ratios. Am J Epidemiol 121:309–323

Gamlem H, Nordstoga K, Glattre E (2008) Canine Neoplasia – Introductory Paper. APMIS 116:5–18

Gavazza A, Presciuttini S, Barale R, et al (2001) Association between canine malignant lymphoma, living in industrial areas, and use of chemicals by dog owners. J Vet Intern Med 15:190–195

Geary RC (1954) The Contiguity Ratio and Statistical Mapping. Incorporated Stat 5:115–127

Gehlke CE, Biehl K (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. J Am Stat Assoc 29:169–170

Getis A, Ord JK (1992) The Analysis of Spatial Association by Use of Distance Statistics. Geogr Anal 24:189–206

Gibbons CL, Mangen M-JJ, Plass D, et al (2014) Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. BMC Public Health 14:

Gilbert A, Chakraborty J (2011) Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. Soc Sci Res 40:273–286

Glass GE (2000) Update: Spatial Aspects of Epidemiology: The Interface with Medical Geography. Epidemiol Rev 22:136–139

Glick B (1979) The Spatial Autocorrelation of Cancer Mortality. Soc Sci Med [Med Geogr] 13:123–130

Glickman LT, Domanski LM, Maguire TG, et al (1983) Mesothelioma in pet dogs associated with exposure of their owners to asbestos. Environ Res 32:305–313

Glickman LT, Schofer FS, McKee LJ, et al (1989) Epidemiologic study of insecticide exposures, obesity, and risk of bladder cancer in household dogs. J Toxicol Envirol Health 28:407–414

Gliklich RE, Dreyer NA, Leavy MB (eds) (2014) Principles of Registry Ethics, Data Ownership, and Privacy. In: Registries for Evaluating Patient Outcomes: A User's Guide. Agency for Healthcare Research and Quality, Rockville, MD, US, pp 145–186

Goldberg J, Gelfand HM, Levy PS (1980) Registry Evaluation Methods: A Review and Case Study. Epidemiol Rev 2:210–220

Goodchild MF (2009) Geographic Information Systems and Science: Today and Tomorrow. Ann GIS 15:3–9

——— (1992) Geographical Data Modeling. Comput Geosci 18:401–408

Gotway CA, Young LJ (2002) Combining Incompatible Spatial Data. J Am Stat Assoc 97:632–648

Graham AJ, Atkinson PM, Danson FM (2004) Spatial Analysis for Epidemiology. Acta Tropica 91:219–225

Greenberg M (1985) Cancer Atlases: Uses and Limitations. Environmentalist 5:187–191

Grüntzig K, Graf R, Boo G, et al (2016) Swiss Canine Cancer Registry 1955–2008: Occurrence of the most common tumour diagnoses and influence of age, breed, body size, sex and neutering status on tumour development. J Comp Pathol 155:156–170

Grüntzig K, Graf R, Hässig M, et al (2015) The Swiss Canine Cancer Registry: A retrospective study on the occurrence of tumours in dogs in Switzerland from 1955 to 2008. J Comp Pathol 152:161–171

Gujarati DN, Porter D (2003) Multicollinearity: What happens if the regressors are correlated. In: Basic Econometrics, 4th ed. McGraw Hill, Boston, MA, US, pp 341–386

Gulliver J, Briggs D, de Hoogh K (2015) Environmental measurement and modelling: introduction and geographical information systems. In: Nieuwenhuijsen M (ed) Exposure Assessment in Environmental Epidemiology. Oxford University Press, Oxford, UK, pp 45–68

Guthrie KA, Sheppard L (2001) Overcoming Biases and Misconceptions in Ecological Studies. J R Stat Soc Ser A Stat Soc 164:141–154

Harbison ML, Godleski JJ (1983) Malignant Mesothelioma in Urban Dogs. Vet Pathol 20:531–540

Hardin JW, Hilbe J (2007) Generalized Linear Models and Extensions, 2nd edn. Stata Press, College Station, TX, US

Haviland A (1875) The Geographical Distribution of Disease in Great Britain, 1st edn. Swan Sonnenschein, London, UK

Hawkins DM (1980) Identification of Outliers. Springer, Dordrecht, The Netherlands

Hayes HM, Hoover R, Tarone RE (1981) Bladder cancer in pet dogs: A sentinel for environmental cancer? Am J Epidemiol 114:229–233

He H, Tang W, Wang W, Crits-Christoph P (2014) Structural Zeroes and Zero-Inflated Models. Shanghai Arch Psychiatry 26:236–242

Heckman JJ (1979) Sample Selection Bias as a Specification Error. Econometrica 47:129–137

——— (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 5:475–492

——— (1974) Shadow Prices, Market Wages, and Labor Supply. Econometrica 42:679–94

Hernán MA, Hernández-Díaz S, Robins JM (2004) A Structural Approach to Selection Bias. Epidemiology 15:615–625

Holt D, Steel DG, Tranmer M, Wrigley N (2010) Aggregation and Ecological Effects in Geographically Based Data. Geog Anal 28:244–261

Hoover R, Mason TJ, McKay FW, Fraumeni JF (1975) Cancer by county: new resource for etiologic clues. Science 189:1005–1007

Hu M-C, Pavlicova M, Nunes EV (2011) Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse 37:367–375

Hyndman RJ, Koehler AB (2006) Measuring Forecast Accuracy. Int J Forecast 22:679–688

Jacquez GM (2004) Current practices in the spatial analysis of cancer: flies in the ointment. Int J Health Geogr 3:22

——— (2000) Spatial analysis in epidemiology: Nascent science or a failure of GIS? J Geograph Syst 2:91–97

Jacquez GM, Jacquez JA (1997) Disease clustering for uncertain locations. In: Lawson AB, Biggeri A, Böhning D, et al. (eds) Disease Mapping and Risk Assessment for Public Health. John Wiley, London, UK, pp 151–168

Jacquez GM, Waller LA (2000) The effect of uncertain locations on disease cluster statistics. In: Mower H, Congalton R (eds) Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing. Ann Arbor Press, Chelsea, MI, US, pp 53–64

Kamel Boulos MN (2004) Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. Int J Health Geogr 3:1

Kanaroglou P, Delmelle E (2016) Spatial Analysis in Health Geography. Routledge, Abingdon, UK

Kimura KC, Carneiro CS, Domenico RM, et al (2015) Cartography of neoplasms in dogs from different regions of the city of São Paulo, SP, Brazil: a survey (2002-2003) of data from the Veterinary Hospital of the School of Veterinary Medicine and Animal Science of the University of São Paulo, Brazil. Braz J Vet Res An Sci 52:257–265

Knox EG (1989) Detection of Clusters. In: Elliott P (ed) Methodology of enquiries into disease clustering, small area health statistics unit. London, UK, pp 17–20

Koch T (2017) Cartographies of Disease: Maps, Mapping, and Medicine. ESRI Press, Redlands, CA, US

Kukull WA, Ganguli M (2012) Generalizability. Neurology 78:1886–1891

Kulldorff M, Athas WF, Feurer EJ, et al (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health 88:1377–1380

Kulldorff M, Feuer EJ, Miller BA, Freedma LS (1997) Breast cancer clusters in the northeast United States: a geographic analysis. Am J Epidemiol 146:161–170

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Statist Med 14:799–810

Lambert D (1992) Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. Technometrics 34:1–14

Lannin DR, Mathews HF, Mitchell J, et al (1998) Influence of socioeconomic and cultural factors on racial differences in late-stage presentation of breast cancer. JAMA 279:1801–1807

Lawson AB (2006) Statistical Methods in Spatial Epidemiology, 2nd edn. John Wiley, Chichester, UK

Lawson AB, Williams FL, Williams F (2001) An introductory guide to disease mapping. John Wiley, Chichester, UK

Legendre P, Legendre LFJ (2012) Numerical Ecology. Elsevier, Amsterdam, The Netherlands

Lengerich EJ, Teclaw RF, Mendlein JM, et al (1992) Pet populations in the catchment area of the Purdue Comparative Oncology Program. J Am Vet Med Assoc 200:51–56

Leung Y, Mei C-L, Zhang W-X (2000) Statistical tests for spatial non-stationarity based on the geographically weighted regression model. Environ Plann A 32:9–32

Levine TR, Hullett CR (2002) Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. Hum Commun Res 28:612–625

Lewis F, Butler A, Gilbert L (2011) A Unified Approach to Model Selection Using the Likelihood Ratio Test. Methods Ecol Evol 2:155–162

Leyk S, Buttenfield BP, Nagle NN, Stum AK (2013) Establishing relationships between parcel data and land cover for demographic small area estimation. Cartogr Geogr Inf Sci 40:305–315

Leyk S, Maclaurin GJ, Hunter LM, et al (2012a) Spatially and temporally varying associations between temporary outmigration and natural resource availability in resource-dependent rural communities in South Africa: A modeling framework. Appl Geogr 34:559–568

Leyk S, Norlund PU, Nuckols JR (2012b) Robust assessment of spatial non-stationarity in model associations related to pediatric mortality due to diarrheal disease in Brazil. Spat Spatiotemporal Epidemiol 3:95–105

Lloyd CD (2014) Exploring Spatial Scale in Geography. John Wiley, Chichester, UK

——— (2010) Local Modelling. In: Local Models for Spatial Analysis, 2nd edn. CRC Press, Boca Raton, FL, US, pp 23–26

Lund EM, Armstrong PJ, Kirk CA, et al (1999) Health status and population characteristics of dogs and cats examined at private veterinary practices in the United States. J Am Vet Med Assoc 214:1336–1341

Lunn D, Spiegelhalter D, Thomas A, Best N (2009) The BUGS Project: Evolution, Critique and Future Directions. Stat Med 28:3049–3067

Maclaurin G, Leyk S, Hunter L (2015) Understanding the combined impacts of aggregation and spatial non-stationarity: The case of migration-environment associations in rural South Africa. Trans GIS 19:877–895

MacMahon B, Pugh TF (1970) Epidemiology: Principles and Methods. Little Brown, Boston, MA, US

MacVean DW, Monlux AW, Anderson PS, et al (1978) Frequency of Canine and Feline Tumors in a Defined Population. Vet Pathol 15:700–715

Marconato L, Leo C, Girelli R, et al (2009) Association between waste management and cancer in companion animals. J Vet Intern Med 23:564–569

Mark D, Turk A (2003) Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. In: Kuhn W, Worboys MF, Timpf S (eds) Spatial Information Theory. Foundations of Geographic Information Science. Springer, Berlin, pp 28–45

Mason TJ (1976) Geographic patterns of cancer risk: A means for identifying possible occupational factors. Ann NY Acad Sci 271:370–376

Mason TJ, McKay FW, Hoover R, et al (1975) Atlas of cancer mortality for US Counties, 1950-1969. U.S. Government Printing Office, Washington, DC, US

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers in Econometrics. Academic Press, New York, NY, US, pp 105–142

McNamee R (2003) Confounding and Confounders. Occup Environ Med 60:227–234

Meade MS, Emch M (2010) Medical Geography. Guilford Press, New York, NY, US

Mennis J (2009) Dasymetric Mapping for Estimating Population in Small Areas. Geogr Compass 3:727–745

—— (2003) Generating surface models of population using dasymetric mapping. Prof Geogr 55:31–42

Mennis J, Hultgren T (2006) Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci 33:179–194

Merlo DF, Rossi L, Pellegrino C, et al (2008) Cancer incidence in pet dogs: findings of the animal tumor registry of Genoa, Italy. J Vet Intern Med 22:976–984

Michell AR (1999) Longevity of British breeds of dog and its relationships with sex, size, cardiovascular variables and disease. Vet Rec 145:625–629

Mohri M, Roark B (2005) Structural Zeros Versus Sampling Zeros. Oregon Health & Science University, Portland, OR, US

Møller Jensen Ø, Carstensen B, Glattre E, et al (1988) Atlas of cancer Incidence in the Nordic Countries, Nordic Cancer Union (The Cancer Societies of Denmark, Finland, Iceland, Norway and Sweden). Puna Musta, Helsinki, Finnland

Mollié A (1990) Représentation géographique des taux de mortalité: modélisation spatiale et méthodes Bayésiennes (unpublished Ph. D. thesis)

Mollié A, Richardson S (1991) Empirical Bayes estimates of cancer mortality rates using spatial models. Stat Med 10:95–112

Moran PAP (1950) Notes on Continuous Stochastic Phenomena. Biometrika 37:17–23

——— (1948) The Interpretation of Statistical Maps. J R Stat Soc Series B Stat Methodol 10:243–251

Morris RD, Munasinghe RL (1993) Aggregation of existing geographic regions to diminish spurious variability of disease rates. Statist Med 12:1915–1929

Nagle NN, Buttenfield BP, Leyk S, Speilman S (2014) Dasymetric Modeling and Uncertainty. Ann Assoc Am Geogr 104:80–95

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc A 231:289–337

NICER (2017) National Statistics on Cancer Incidence. http://www.nicer.org/en/statistics-atlas/cancer-incidence. Accessed 31 Dec 2017

Nødtvedt A, Berke O, Bonnett BN, Brønden L (2011) Current status of canine cancer registration – Report from an international workshop. Vet Comp Oncol 10:95–101

O'Brien DJ, Kaneene JB, Getis A, et al (2000) Spatial and temporal comparison of selected cancers in dogs and humans, Michigan, USA, 1964–1994. Prev Vet Med 47:187–204

Olson KL, Grannis SJ, Mandl KD (2006) Privacy Protection Versus Cluster Detection in Spatial Epidemiology. Am J Public Health 96:2002–2008

O'Neill DG, Church DB, McGreevy PD, et al (2014) Approaches to Canine Health Surveillance. Canine Genet Epidemiol 1:1–13

Openshaw S (1984) The Modifiable Areal Unit Problem — Concepts and Techniques in Modern Geography. Geo Books, Norwich, UK

Openshaw S, Charlton ME, Wymer C, Craft A (1987) Geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inf Sci 1:335–358

Openshaw S, Craft AW, Charlton ME, Birch JM (1988) Investigation of leukemia clusters by the use of a Geographical Analysis Machine. Bailliere Tindall, London, UK

Openshaw S, Taylor P (1979) A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In: Wrigley N (ed) Statistical Methods in the Spatial Sciences. Routledge and Kegan Paul, London, UK, pp 127–144

Owen LN (1979) A Comparative Study of Canine and Human Breast Cancer. Invest Cell Pathol 2:257–275

Ozonoff A, Jeffery C, Manjourides J, et al (2007) Effect of spatial resolution on cluster detection: A simulation study. Int J Health Geogr 6:

Páez A, Delmelle E, Kanaroglou P (eds) (2015) Spatial Analysis in Health Geography. Routledge, New York, NY, US

Parkin DM (2008) The Role of Cancer Registries in Cancer Control. Int J Clin Oncol 13:102–111

——— (2006) The Evolution of the Population-Based Cancer Registry. Nat Rev Cancer 6:603–612

Pastor M, Chalvet-Monfray K, Marchal T, et al (2009) Genetic and environmental risk indicators in canine non-Hodgkin's lymphomas: Breed associations and geographic distribution of 608 cases diagnosed throughout France over 1 year. J Vet Intern Med 23:301–310

Pearce N (2012) Classification of Epidemiological Study Designs. Int J Epidemiol 41:393–397

Perloff JD, Kletke PR, Fossett JW, Banks S (1997) Medicaid participation Among Urban Primary Care Physicians. Med Care 35:142–157

Petrov A (2012) One Hundred Years of Dasymetric Mapping: Back to the Origin. Cartogr J 49:256–264

Piantadosi S, Byar DP, Green SB (1988) The Ecological Fallacy. Am J Epidemiol 127:893–904

Picard RR, Cook RD (1984) Cross-Validation of Regression Models. J Am Stat Assoc 79:575–583

Pincus T, Esther R, DeWalt DA, Callahan LF (1998) Social conditions and self-management are more powerful determinants of health than access to care. Ann Intern Med 129:406–411

Pinho SS, Carvalho S, Cabral J, et al (2012) Canine Tumors: A Spontaneous Animal Model of Human Carcinogenesis. Transl Res 159:165–172

Ponce F, Marchal T, Magnol JP, et al (2010) A morphological study of 608 cases of canine malignant lymphoma in France with a focus on comparative similarities between canine and human lymphoma morphology. Vet Pathol 47:414–433

Pospischil A, Grüntzig K, Graf R, et al (2015) One Medicine – One Oncology – Incidence and Geographical Distribution of Tumors in Dogs and Cats in Switzerland from 1955-2008. Proceedings of the GRF One Health Summit 2015 108–111

Pospischil A, Hässig M, Vogel R, et al (2013) Hundepopulation und Hunderassen in der Schweiz von 1955 bis 2008. Schweiz Arch für Tierheilkd 155:219–228

Potosky AL, Breen N, Graubard BI, Parsons PE (1998) The association between health care coverage and the use of cancer screening tests: Results from the 1992 national health interview survey. Med Care 36:257–270

Preisser JS, Stamm JW, Long DL, Kincade ME (2012) Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Res 46:413–423

Proschowsky HF, Rugbjerg H, Ersbøll AK (2003) Mortality of Purebred and Mixed-breed Dogs in Denmark. Prev Vet Med 58:63–74

Rabinowitz P, Scotch M, Conti L (2009) Human and Animal Sentinels for Shared Health Risks. Vet Ital 45:23–24

Ragland W, Gorham JR (1967) Tonsillar Carcinoma in Rural Dogs. Nature 214:925–926

Raubal M, Jacquez G, Wilson J, Kuhn W (2013) Synthesizing Population, Health, and Place. JOSIS 1–6

Reibel M, Bufalino ME (2005) Street-weighted interpolation techniques for demographic count estimation in incompatible zone system. Environ Plann A 37:127–139

Reif J, Rhodes WH, Cohen D (1970) Canine Pulmonary Disease and the Urban Environment. Arch Environ Occup Health 20:676–683

Reif JS (2011) Animal Sentinels for Environmental and Public Health. Public Health Rep 126:50–57

Reif JS, Bruns C, Lower KS (1998) Cancer of the nasal cavity and paranasal sinuses and exposure to environmental tobacco smoke in pet dogs. Am J Epidemiol 147:488–492

Reif JS, Cohen D (1970) Retrospective radiographic analysis of pulmonary disease in rural and urban dogs. Arch Environ Occup Health 20:684–689

Reif JS, Lower KS, Ogilvie GK (1995) Residential exposure to magnetic fields and risk of canine lymphoma. Am J Epidemiol 141:352–359

Rezaeian M, Dunn G, St Leger S, Appleby L (2007) Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. J Epidemiol Community Health 61:98–102

Richardson S (1996) Statistical Methods for Geographical Correlation Studies. In: Elliott P, Cuzick J (eds) Geographical and Environmental Epidemiology – Methods for Small-Area Studies. Oxford University Press, Oxford, UK, pp 181–204

Richardson S, Montfort C (2000) Ecological Correlation Studies. In: Elliott P, Wakefield J, Best N, Briggs D (eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, UK

Robles SC, Marrett LD, Clarke EA, Risch HA (1988) An application of capture-recapture methods to the estimation of completeness of cancer registration. J Clin Epidemiol 41:495–501

Roquette R, Painho M, Nunes B (2017) Spatial epidemiology of cancer: a review of data sources, methods and risk factors. Geospat Health 12:

Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D (1999) Spatial Autocorrelation of Cancer in Western Europe. Eur J Epidemiol 15:15–22

Roth RE, Woodruff AW, Johnson ZF (2010) Value-by-alpha maps: An alternative technique to the cartogram. Cartogr J 47:130–140

Rowell JL, McCarthy DO, Alvarez CE (2011) Dog Models of Naturally Occurring Cancer. Trends Mol Med 17:380–388

Schmidt PL (2009) Companion Animals as Sentinels for Public Health. Vet Clin North Am Small Anim Pract 39:241–250

Schneider R (1970) Comparison of age, sex, and incidence rates in human and canine breast cancer. Cancer 26:419–426

Schneider R, Dorn CR, Klauber MR (1968) Cancer in Households. A Human-Canine Retrospective Study. J Nat Cancer Inst 41:1285–1292

Schouten LJ, Straatman H, Kiemeney LA, et al (1994) The capture-recapture method for estimation of cancer registry completeness: a useful tool? Int J Epidemiol 23:1111–1116

Scotch M, Odofin L, Rabinowitz P (2009) Linkages Between Animal and Human Health Sentinel Data. BMC Vet Res 5:

Sethi D, Wheeler J, Rodrigues LC, et al (1999) Investigation of under-ascertainment in epidemiological studies based in general practice. Int J Epidemiol 28:106–112

SFOT (2017) Federal Office of Topography - Swisstopo. http://www.swisstopo.admin.ch. Accessed 31 Dec 2017

SFSO (2017) Swiss Federal Statistical Office. http://www.bfs.admin.ch. Accessed 31 Dec 2016

SFTA (2017) Swiss Federal Tax Administration. http://www.estv.admin.ch. Accessed 31 Dec 2017

Shields PM, Rangarajan N (2013) A playbook for research methods: Integrating conceptual frameworks and project management. New Forums Press, Stillwater, OK, US

Snee RD (1977) Validation of Regression Models: Methods and Examples. Technometrics 19:415–428

Snow J (1855) On the Mode of Communication of Cholera. John Churchill, London, UK

Spearman C (1904) The Proof and Measurement of Association between Two Things. Am J Psychol 15:72–101

Sporn MB (1996) The War on Cancer. Lancet 347:1377–1381

St. Sauver JL, Grossardt BR, Leibson CL, et al (2012) Generalizability of epidemiological findings and public health decisions: An illustration from the Rochester epidemiology project. Mayo Clin Proc 87:151–160

Steyerberg EW, Harrell Jr FE, Borsboom GJJM, et al (2001) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54:774–781

Stock P (1939) Distribution in England and Wales of Cancer of Various Organs. British Empire Cancer Campaign, London, UK

—— (1937) Distribution in England and Wales of Cancer of Various Organs. British Empire Cancer Campaign, London, UK

—— (1936) Distribution in England and Wales of Cancer of Various Organs. British Cancer Campaign, London, UK

—— (1928) On the evidence for a regional distribution of cancer prevalence in England and Wales. British Empire Cancer Campaign, London, UK

Stokes CS, Brace KD (1988) Agricultural chemical use and cancer mortality in selected rural counties in the U.S.A. J Rural Stud 4:239–247

Strafuss AC (1976) Sebaceous Gland Adenomas in Dogs. J Am Vet Med Assoc 169:640–642

Swerdlow A, Silva IDS (eds) (1992) Atlas of Cancer Incidence in England and Wales, 1968-85. Oxford University Press, Oxford, UK

Swisscom Ltd. (2017) The Official Phonebook and Yellow Pages of Switzerland. http://www.local.ch. Accessed 31 Dec 2017

Tapp AF (2010) Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources. Cartogr Geogr Inf Sci 37:215–228

Tate N, Atkinson PM (2001) Changing the Scale of Measurement. In: Modelling Scale in Geographical Information Science. John Wiley, Hoboken, NJ, US, pp 159–260

Tedardi MV, Veneziano DB, Kimura KC, et al (2015) Sao Paulo Animal Cancer Registry, the First in Latin America. Vet Comp Oncol 13:154–155

Teppo L, Pukkala E, Lehtonen M (1994) Data quality and quality control of a population-based cancer registry: experience in Finland. Acta Oncologica 33:365–369

The Canary Database (2017) Linkage to Human Health. https://canarydatabase.org/about/linkage. Accessed 31 Dec 2017

Thompson SK (2012) Sampling, 3rd edn. John Wiley, Hoboken, NJ, US

Thygesen LC, Ersbøll AK (2014) When the entire population is the sample: strengths and limitations in register-based epidemiology. Eur J Epidemiol 29:551–558

Tiefelsdorf M (2006) Modelling spatial processes: The identification and analysis of spatial relationships in regression residuals by means of Moran's I. Springer, New York, NY, US

Tobler WR (1989) Frame Independent Spatial Analysis. In: Goodchild MF, Gopal S (eds) The Accuracy of Spatial Databases. CRC Press, Boca Raton, FL, US, pp 75–79

——— (1979) Smooth Pycnophylactic Interpolation for Geographical Regions. J Am Stat Assoc 74:519–530

Torre LA, Siegel RL, Ward EM, Jemal A (2016) Global Cancer Incidence and Mortality Rates and Trends — An Update. Cancer Epidemiol Biomarkers Prev 25:16–27

Tsai P-J, Perng C-H (2011) Spatial autocorrelation analysis of 13 leading malignant neoplasms in Taiwan: a comparison between the 1995-1998 and 2005-2008 periods. Health 3:712

Vascellari M, Baioni E, Ru G, et al (2009) Animal tumour registry of two provinces in northern Italy: Incidence of spontaneous tumours in dogs and cats. BMC Vet Res 5:

Vega Orozco CD, Golay J, Kanevski M (2015) Multifractal Portrayal of the Swiss Population. Cybergeo

Vineis P, Wild CP (2014) Global Cancer Patterns: Causes and Prevention. Lancet 383:549–557

Vuong QH (1989) Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57:307–333

Wakefield J, Kelsall J, Morris S (2000) Clustering, Cluster Detection, and Spatial Variation in Risk. In: Elliott P, Wakefield J, Best N, Briggs D (eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, UK, pp 128–152

Wall MM (2004) A close look at the spatial structure implied by the CAR and SAR models. J Stat Plan Inference 121:311–324

Waller LA, Gotway CA (2004) Applied Spatial Statistics for Public Health Data. John Wiley, Hoboken, NJ, US

Waller LA, Hill EG, Rudd RA (2006) The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Statist Med 25:853–865

Waller LA, Turnbull BW (1993) The Effects of Scale on Tests for Disease Clustering. Statist Med 12:1869–1884

Walter SD (2000) Disease Mapping: A Historical Perspective. In: Elliott P, Wakefield J, Best N, Briggs D (eds) Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford, UK, pp 332–239

Walter SD, Birnie SE (1991) Mapping Mortality and Morbidity Patterns: An International Comparison. Int J Epidemiol 20:678–689

Walter SD, Birnie SE, Marrett LD, et al (1994) The Geographic Variation of Cancer Incidence in Ontario. Am J Public Health 84:367–376

Wan X, Wang W, Liu J, Tong T (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol 14:135

Ward MH, Wartenberg D (2006) Invited commentary: On the road to improved exposure assessment using geographic information systems. Am J Epidemiol 164:208–211

Wartenberg D (1999) Using disease-cluster and small-area analyses to study environmental justice. In: Toward Environmental Justice: Research, Education, and Health Policy Needs, National Academies Press. Washington, DC, US, pp 79–102

White R (1972) Probability Maps of Leukemia Mortality in England and Wales. In: McGlashan N (ed) Medical Geography – Techniques and Field Studies. Methuen, London, UK, pp 171–85

Whittle P (1954) On Stationary Processes in the Plane. Biometrika 41:434–449

WHO (2015) International Classification of Diseases for Oncology – 3rd Edition (ICD-O-3). http://www.who.int/classifications/icd/adaptations/oncology. Accessed 31 Dec 2017

WHO (2017) Cancer Fact Sheet – February 2017. http://www.who.int/mediacentre/factsheets/fs297. Accessed 31 Dec 2017

Williamson DF, Parker RA, Kendrick JS (1989) The Box Plot: A Simple Visual Method to Interpret Data. Ann Intern Med 110:916–921

Willmott CJ (1981) On the Validation of Models. Phys Geogr 2:184–194

Wilson P (2015) The Misuse of the Vuong Test for Non-Nested Models to Test for Zero-inflation. Econ Lett 127:51–53

Wirth KE, Tchetgen EJ (2014) Accounting for selection bias in association studies with complex survey data. Epidemiology 25:444–453

Wong D (2009) The Modifiable Areal Unit Problem (MAUP). SAGE Publications: London, UK

Woodward M (2013) Epidemiology: Study Design and Data Analysis, 3rd edn. CRC Press, Boca Raton, FL, US

Wright JK (1936) A method of mapping densities of population: with Cape Cod as an example. Geogr Rev 26:103–110

Zandbergen PA (2011) Dasymetric mapping using high resolution address point datasets. Trans GIS 15:5–27

Zeileis A, Kleiber C, Jackman S (2008) Regression Models for Count Data in R. J Stat Softw 27:1–25

Zhou H-B, Liu S-Y, Lei L, et al (2015) Spatio-temporal analysis of female breast cancer incidence in Shenzhen, 2007–2012. Chin J Cancer 34:13

Zoraghein H, Leyk S, Ruther M, Buttenfield BP (2016) Exploiting temporal information in parcel data to refine small area population estimates. Comput Environ Urban Syst 58:19–28

7. ANNEXES

7.1. REFERENCE MAPS

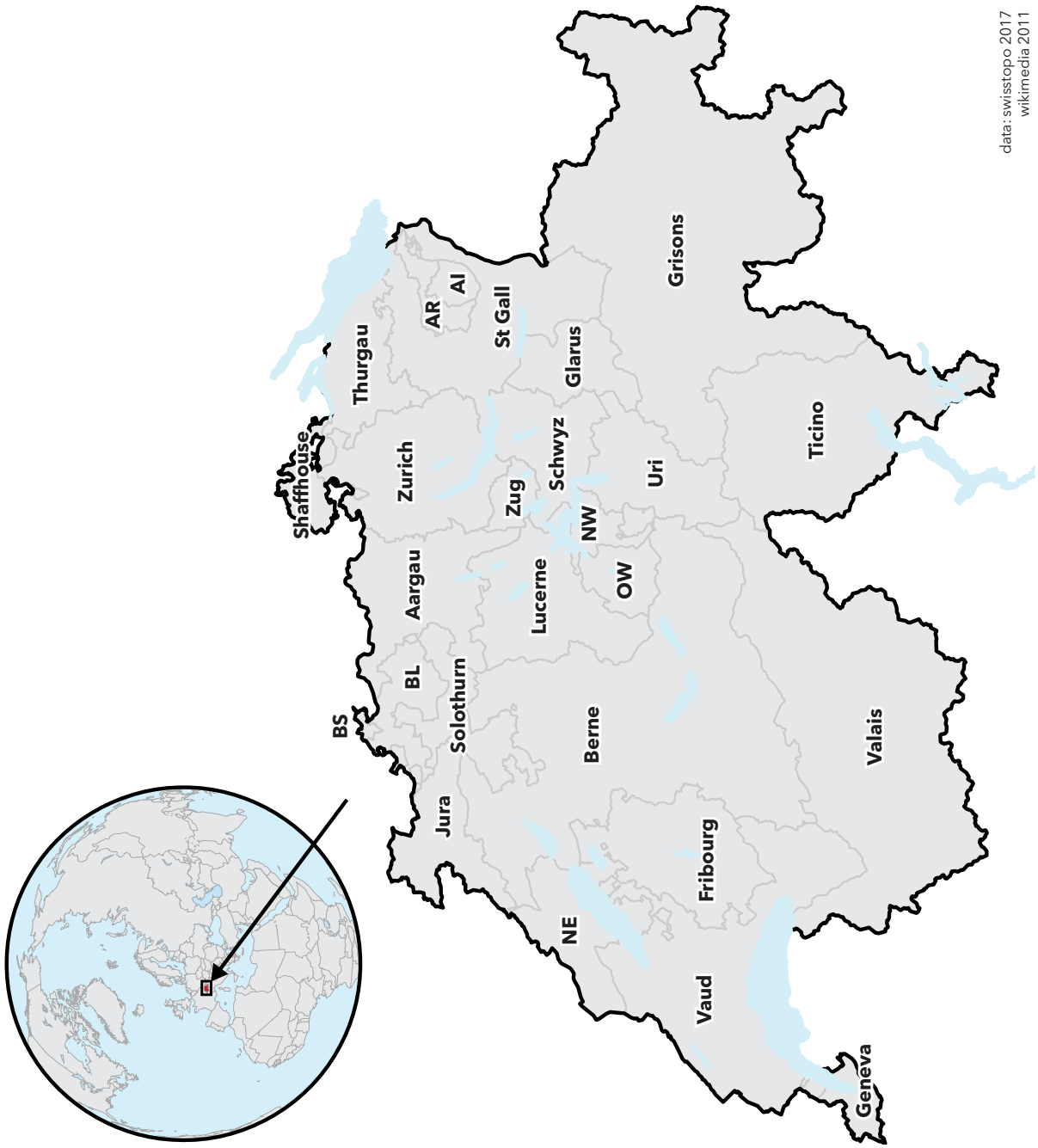

Figure 16. Location map of Switzerland and boundaries of the Swiss cantons. The name of following cantons is abbreviated – Appenzell Innerrhoden (AI), Appenzell Ausserrhoden (AR), Basle-Country (BL), Basle-City (BS), Neuchâtel (NE), Nidwalden (NW), and Obwalden (OW).

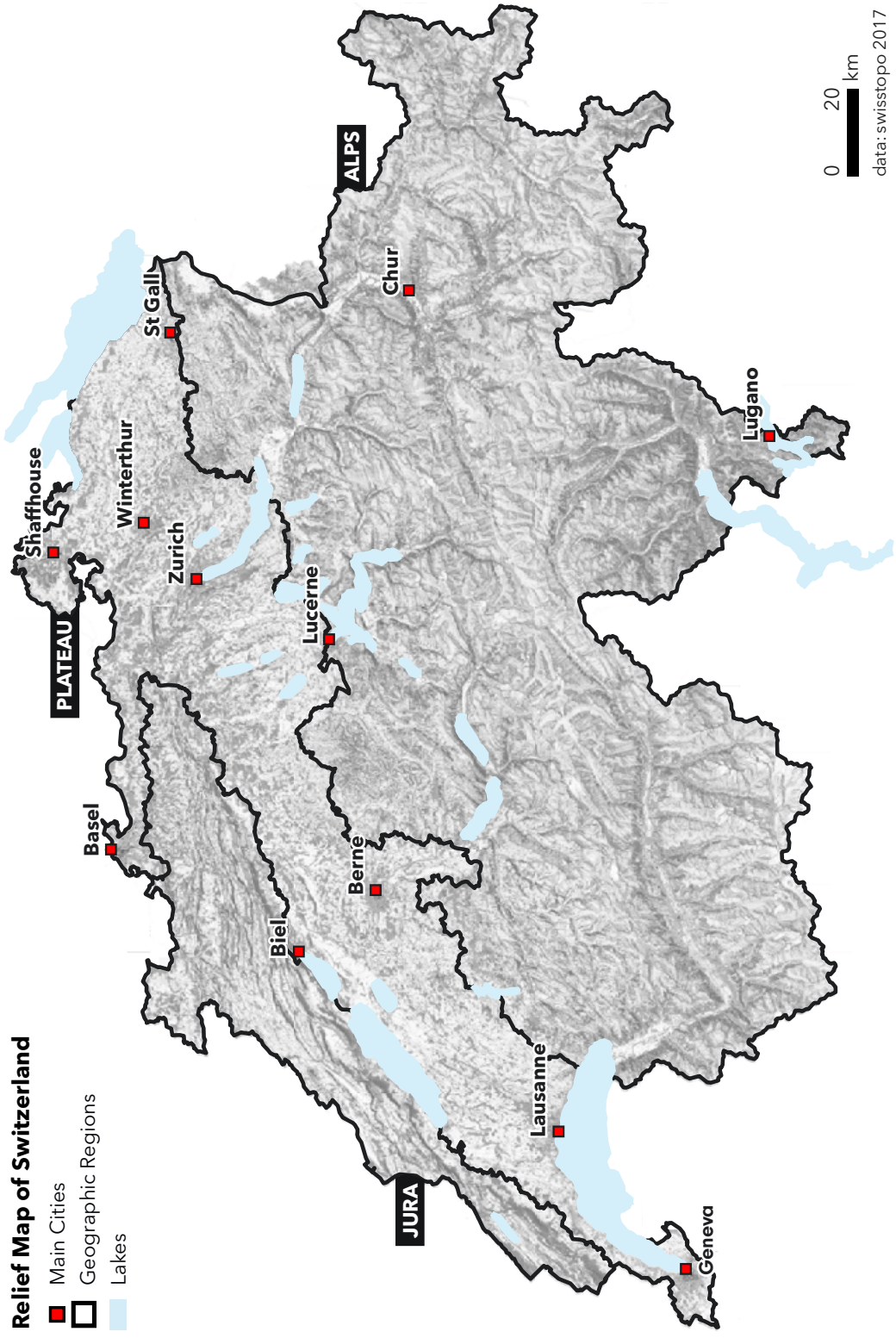

Figure 17. Relief map of Switzerland. The main cities and geographic regions are also overlaid.

7.2. RELEVANT KNOWLEDGE AREAS OF GISCIENCE

KA 1 – ANALYTICS AND MODELING (AM)

adapted from DiBiase et al. (2017)

Basic Spatial Operations	**Advanced Spatial Analysis**	**Surface Analysis**
(04) Buffers	(53) Analytical procedures	(15) Surface derivatives
(05) Overlay	(07) Point pattern analysis	(16) Interpolation methods
(06) Neighborhoods	(09) Cluster analysis	(17) Intervisibility
(07) Map algebra	(19) Exploratory data analysis	(18) Cost surfaces
Spatial Modeling	(11) Multi-dim attributes	**Network Analysis**
(12) Cartographic modeling	(13) Multi-criteria evaluation	(40) Least-cost path analysis
(50) Components of models	(89) Weighting schemes	(41) Flow modeling
(54) Scientific models with GIS	(10) Spatial interaction	(42) Classic Transport Problem
( ) Mathematical models	(21) Spatial weights matrix	(43) Classic network problems
(14) Spatial process models	(67) Space-scale algorithms	(44) Modeling Accessibility
(49) Represent info & process	**Space-Time Anal & Model**	**Data Mining**
(55) Analysis & design	(90) Movement analysis	(36) Data mining approaches
Data Manipulation	( ) Time geography	(37) Knowledge discovery
(62) Point, line & area gen	**Spatial Statistics**	(38) Pattern recognition
(61) Coordinate transform	(22) Global spatial association	(65) Geospatial data class
(57) Data conversion	(23) Local spatial association	(66) MLFF neural networks
(56) Impacts of transformation	(26) Spatial sampling	(68) Rule learning
(60) Raster resampling	(20) Stochastic processes	**Spatial Simulation**
(59) VtR & RtV conversions	(24) Outliers	(84) Simulation modeling
Errors and Uncertainty	(25) Bayesian methods	(69) Cellular Automata
(87) Currency, source & scale	(27) Semi-variogram constr	(76) Simulated annealing
(86) Theory of error propag	(28) Semi-variogram model	(79) Agent-based models
(85) Propagation of error	(29) Kriging methods	(81) Adaptive agents
(88) Fuzzy aggregation	(31) Spatial econometrics	(82) Microsim & calibration
	(32) SAR	**Spatial Optimization**
	(33) Spatial filtering	(46) Loc-allocation modeling
	(08) Kernels and density estim	(73) Greedy heuristics
	(34) Spatial expansion & GWR	(74) Interchange heuristics
	(47) Spatial distribution	(18) Genetic algorithms
	(48) Math models uncertainty	
	(63) Non-linearity relationship and non-Gaussian distribution	
	(75) Interchange with prob	

KA 2 – CARTOGRAPHY AND VISUALIZATION (CV)

adapted from DiBiase et al. (2017)

History & Trends	**Map Design Techniques**	**Interactive Design Tech**
(**01**) Cartography & Science	(**11**) Common Thematic Maps	(**13**) UI/UX Design
(**02**) Cartography & Tech	(**12**) Bivariate & Multivar Maps	(**15**) Web Mapping
( ) Cartography & Power	(**17**) Mapping Time	(**16**) Virtual & Immersive Envir
( ) Cartography & Education	(**18**) Mapping Uncertainty	(**19**) Big Data Visualization
( ) Cartography & Art	(**14**) Terrain Representation	( ) Mobile Mapping & Design
Data Considerations	(**32**) Cartograms	( ) Usability Engineering
(**03**) Vector Formats & Sources	( ) Icon Design	( ) Basemaps
(**20**) Raster Formats & Sources	( ) Narrative & Storytelling	( ) Geovisualization
(**25**) Metadata, Quality & Unc	( ) Flow Maps	( ) Geocollaboration
Map Design Fundamentals	**Map Use**	(**36**) Geovisual Analytics
(**04**) Scale & Generalization	(**21**) Map Reading	
(**05**) Statistical Mapping	(**22**) Map Interpretation	
( ) Geodesy, Coord & Proj	(**23**) Map Analysis	
(**07**) Visual Hier, Layout & Elem	(**24**) User-Centered Design	
(**08**) Symboliz & Visual Var	( ) Political Economy of Map	
(**09**) Color Theory	( ) Map Critique	
(**10**) Typography		
( ) Aesthetics & Design		
( ) Map Production & Manag		

KA 3 – GISCIENCE&TECHNOLOGY AND SOCIETY (GS)

adapted from DiBiase et al. (2017)

Cognitive and Social Found	**Governance and Agency**	**GI as Property**
(**17**) Common-sense Geo	(**11**) Prof & Pract Eth of GIS&T	(**07**) Property Regimes
(**18**) Cultural Influences	(**12**) Cod of Eth for GS Prof	(**08**) Mech of Control of GI
(**19**) Political Influences	(**24**) Citizen Sci with GIS&T	(**09**) Enforcing Control of GI
( ) Alt Representations	( ) Spatial Decision Support	( ) Geopiracy
Law, Regulation, and Policy	( ) Marginal Societies	**Critical Perspectives**
(**01**) The Legal Regime	(**21**) Sec & OA to GS Data	(**13**) Epistemological Crit
(**03**) Liability	(**06**) Public Participation GIS	(**14**) Ethical Critiques
(**02**) Contract Law	(**22**) Implic of Dist GIS&T	(**15**) Feminist Critiques
(**04**) Location Privacy	(**20**) Aggregation of Spat Ent	(**16**) Social Critiques
(**23**) Sharing Geospatial Info		(**10**) Data Access, Sec & Priv

7.3. CURRICULUM VITAE

BOO Gianluca
Born March 2, 1983 in Faido TI, Switzerland
Native of Bodio TI, Switzerland

EDUCATION

2014-2018 **Dr.sc.nat. in Geography**
Department of Geography, University of Zurich, Switzerland

Title of the thesis: Dogs as sentinels for environmental cancers – Addressing the challenges of spatial epidemiology

Committee: Prof. Dr. Sara I. Fabrikant, Prof. Dr. Andreas Pospischil, Prof. Dr. Kay W. Axhausen, Prof. Dr. Robert Weibel, and Prof. Dr. Stefan Leyk

Subject area: GIScience

2007-2009 **MSc in Geography**
Faculty of Geosciences and Environment, University of Lausanne, Switzerland

Title of the thesis: L'urbanisme à l'épreuve de la durabilité. Projet pour l'écoquartier des Plaines-du-Loup.

Subject area: Urban Studies

2003-2006 **BSc in Geography**
Faculty of Geosciences and Environment, University of Lausanne, Switzerland

Subject area: Human Geography

1998-2003 **Maturità Liceale (Baccalaureate)**
Liceo Cantonale di Bellinzona, Switzerland

Subject area: Physics and Chemistry

EMPLOYMENT AT THE UNIVERSITY OF ZURICH

2014-2018 **Research and Teaching Assistant**
Department of Geography, University of Zurich, Switzerland

7.4. LIST OF PUBLICATIONS

This thesis presents only part of the work carried out within the "One Medicine – One Oncology" research project. Additional featured publications are reported hereafter and marked with a star (*****).

PEER-REVIEWED PUBLICATIONS

Boo G, Leyk S, Fabrikant SI, Graf R, Pospischil A (2018) Assessing uncertainty in canine cancer data sources through dasymetric refinement. ISPRS Int J Geo-Inf [submitted]

Boo G, Leyk S, Brunsdon C, Graf R, Pospischil A, Fabrikant SI (2018) The importance of regional models in assessing canine cancer incidences in Switzerland. PLOS ONE

Boo G, Leyk S, Fabrikant SI, Pospischil A, Graf R (2017) Assessing effects of structural zeros on models of canine cancer incidence: a case study of the Swiss Canine Cancer Registry. Geospat Health 12:121–129

Grüntzig K, Graf R, **Boo G**, Guscetti F, Hässig M, Axhausen KW, Fabrikant SI, Welle M, Meier D, Folkers G, Pospischil A (2016) Swiss Canine Cancer Registry 1955–2008: Occurrence of the most common tumour diagnoses and influence of age, breed, body size, sex and neutering status on tumour development. J Comp Pathol 155:156–170 *****

Grüntzig K, Graf R, Hässig M, Welle M, Meier D, Lott D, Erni D, Schenker NS, Guscetti F, **Boo G**, Axhausen KW, Fabrikant SI, Folkers G, Pospischil A (2015) The Swiss Canine Cancer Registry: A retrospective study on the occurrence of tumours in dogs in Switzerland from 1955 to 2008. J Comp Pathol 154:161–171 *****

Pospischil A, Grüntzig K, Graf R, **Boo G**, Hässig M, Welle M (2015) Krebsregister für Hunde und Katzen in der Schweiz (1955-2008). Pneumologie 69:5–10 *****

NON-REFEREED PUBLICATIONS

Boo G, Leyk S, Fabrikant SI (2016) Dasymetric mapping for an improved modeling of diseases. Geomatik Schweiz/Géomatique Suisse, Sonderheft International Map Year, 115:100–101 *****

PEER-REVIEWED CONFERENCE PROCEEDINGS

Boo G, Leyk S, Fabrikant SI, Pospischil A (2016) A regional approach for modeling dog cancer incidences with regard to different reporting practices. Ninth International Conference on GIScience – Short Paper Proceedings, Montreal, Canada 29–33 [best poster award] *****

Boo G, Fabrikant SI, Leyk S (2015) A novel approach to veterinary spatial epidemiology: dasymetric refinement of the Swiss Dog Tumor Registry data. ISPRS Annals II–3/W5 263–269 *****

Pospischil A, Grüntzig K, Graf R, **Boo G**, Folkers G, Otto V, Fabrikant SI (2015) One Medicine – One Oncology – Incidence and geographical distribution of tumors in dogs and cats in Switzerland from 1955-2008. GRF One Health Summit 2015 – Conference Proceedings, Davos, Switzerland, 108–111 *****

NON-REFEREED CONFERENCE PROCEEDINGS

Boo G, Leyk S, Brunsdon C, Pospischil A, Fabrikant SI (2017) Spatial non-stationarity and geographic scale in models of canine cancer incidence. Proceedings of the GEOMED 2017 Conference. Porto, Portugal *****

Boo G (2016) Spatial Distribution of Dog Cancer in Switzerland: A Case Study of Skin Tumors During the Period 2008– 2013. One Medicine – One Oncology: Animal Cancer Registry Symposium. Zurich, Switzerland *****

Boo G, Leyk S, Pospischil A, Fabrikant SI (2015) An innovative geographical approach to spatial epidemiology – dasymetric refinement of dog tumor incidence location data in Switzerland. Proceedings of the GEOMED 2015 Conference. Florence, Italy [best poster award] *****

Boo G, Fabrikant SI, Leyk S (2015) An Innovative Approach to Spatial Epidemiology: Dasymetric Refinement of Disease Location Data. Spatial Information for Human Health (SPATIAL 2015), Santa Barbara (CA), US *****

BOOK CHAPTERS

Pospischil A, Graf R, Grüntzig R, **Boo G** (2016) Spontaneous Animal Tumor Models. In Schubiger PA, Martic-Kehl MI (eds) Animal Models for Human Cancer: Discovery and Development of Novel Therapeutics, Methods and Principles. John Wiley & Sons, Hoboken, NJ, US, pp 129-152 *****